DOCUMENT RESUME

ED 214 944 TM 820 026

AUTHOR	Wilcox, Rand R.
TITLE	Test Design Project: Studies in Test Adequacy. Annual Report.
INSTITUTION	California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY	National Inst. of Education (ED), Washington, D.C.
PUB DATE	Nov 81
GRANT	NIE-G-80-0112
NOTE	289p.; For related documents see ED 211 592 and ED 212 650.
EDRS PRICE	MF01/PC12 Plus Postage.
DESCRIPTORS	Achievement Tests; Criterion Referenced Tests; Guessing (Tests); *Mathematical Models; *Multiple Choice Tests; Scoring Formulas; Testing Problems; Test Items; *Test Reliability; Test Theory
IDENTIFIERS	*Answer Until Correct; *Distractors (Tests)

ABSTRACT

These studies in test adequacy focus on two problems: procedures for estimating reliability, and techniques for identifying ineffective distractors. Fourteen papers are presented on recent advances in measuring achievement (a response to Molenaar); "an extension of the Dirichlet-multinomial model that allows true score and guessing to be correlated"; results on an answer-until-correct scoring procedure; the k out of n reliability of a test, and an exact test for random guessing; "determining the length of multiple choice criterion-referenced tests when an answer-until-correct scoring procedure is used"; "a closed sequential procedure for comparing the binomial distribution to a standard"; "a closed sequential procedure for answer-until-correct tests"; "approximating the probability of identifying the most effective treatment for the case of normal distributions having unknown and unequal variances"; estimating the reliability of a mastery test with the beta-binomial model; "analyzing the distractors of multiple choice test items or partitioning multinomial cell probabilities with respect to a standard"; "solving measurement problems with an answer-until-correct procedure"; and "a polarization test for making inferences about the entropy of multiple-choice test items." (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Center for the Study of Evaluation

UCLA Graduate School of Education
Los Angeles, California 90024

ED214944

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

CSE
CSECSE
CSECSECSE
CSECSE
CSE

TM 820 026

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J.C. Beer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

DELIVERABLE - November 1981

TEST DESIGN PROJECT: STUDIES IN TEST ADEQUACY

ANNUAL REPORT

Rand Wilcox, Study Director

Grant Number
NIE-G-80-0112
P-3

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pusuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position of policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Table of Contents

INTRODUCTION

(A) METHODS AND RECENT ADVANCES IN MEASURING ACHIEVEMENT: A RESPONSE TO MOLENAAR

(B) AN EXTENSION OF THE DIRICHLET-MULTINOMIAL MODEL THAT ALLOWS TRUE SCORE AND GUESSING TO BE CORRELATED

(C) SOME EMPIRICAL AND THEORETICAL RESULTS ON AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE

(D) SOME NEW RESULTS ON AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE

(E) USING RESULTS ON k OUT OF n SYSTEM RELIABILITY TO STUDY AND CHARACTERIZE TESTS

(F) BOUNDS ON THE K OUT OF N RELIABILITY OF A TEST, AND AN EXACT TEST FOR RANDOM GUESSING

(G) DETERMINING THE LENGTH OF MULTIPLE CHOICE CRITERION-REFERENCED TESTS WHEN AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE IS USED

(H) A CLOSED SEQUENTIAL PROCEDURE FOR COMPARING THE BINOMIAL DISTRIBUTION TO A STANDARD

(I) A CLOSED SEQUENTIAL PROCEDURE FOR ANSWER-UNTIL-CORRECT TESTS

(J) APPROXIMATING THE PROBABILITY OF IDENTIFYING THE MOST EFFECTIVE TREATMENT FOR THE CASE OF NORMAL DISTRIBUTIONS HAVING UNKNOWN AND UNEQUAL VARIANCES

(K) A CAUTIONARY NOTE ON ESTIMATING THE RELIABILITY OF A MASTERY TEST WITH THE BETA-BINOMIAL MODEL

(L) ANALYZING THE DISTRACTORS OF MULTIPLE-CHOICE TEST ITEMS OR PARTITIONING MULTINOMIAL CELL PROBABILITIES WITH RESPECT TO A STANDARD

(M) SOLVING MEASUREMENT PROBLEMS WITH AN ANSWER-UNTIL-CORRECT PROCEDURE

(N) A POLARIZATION TEST FOR MAKING INFERENCES ABOUT THE ENTROPY OF MULTIPLE-CHOICE TEST ITEMS

INTRODUCTION

CSE Studies in Test Adequacy focused on two theoretical problems during FY1981: 1) procedures for estimating reliability and 2) improved techniques for identifying ineffective distractors. Applications of these techniques were also to be demonstrated in the analysis of multiple choice tests. As any psychometrician will agree, the areas of reliability and identifying distractors are intimately related. Progress on one area is likely to influence thought on the other. Although, for the purposes of this report, the progress of research is divided into two discrete sections, in fact, selected papers integrate findings in the two areas. In the August, 1980 plan for the Test Design Project, it was proposed that these analyses consider data from the study of "Literacy Assessment in a School District Context." However, at the request of the NIE, the latter study was deleted from CSE's scope of work; as a result, empirical analyses trying out newly proposed solutions used available data.

Work in Studies of Test Adequacy proceeded faster than anticipated. Initial solutions required little revision, and an important new technique proved very valuable in addressing several test adequacy problems. As a result, more work than anticipated was completed, and an additional aspect of reliability, test length, was also addressed, although not required by the scope of work.

The accomplishments for the year are briefly described below, including work directly related to each problem area and the extension of the developed solutions to other contexts.

Note on Methodology

The methodology used in testing the solutions in the problem areas of reliability and identifying distractors is a mathematical one. In general, it depends upon the positing of a "lemma", a mathematical statement presumed to solve a given problem, and then testing mathematically the quality of the solution. In the text of this report, as the process of exploring potential solutions is traced, "advantages" and "limitations" are noted but not fully described. These terms are used in the mathematical sense and are not matters of personal preference. Advantages (or limitations) of potential solutions are demonstrated mathematically in the coordinate, referenced research papers prepared in this project and are obvious by inspection of the equations.

The estimation of reliability

The problem with estimating the reliability of tests is that the usual and customary estimation procedures either ignore the problem of guessing altogether or make clearly inappropriate assumptions about how guessing affects the data. One frequent assumption is that guessing is completely random. At the beginning of the year it seemed that existing latent structure models might provide a solution to the guessing issue. However, the obvious problem with this solution is that the required model deamnds mathematical assumptions that are frequently impossible to meet. Elaboration of this view is contained in the report entitled "Methods and Recent Advances on in Measuring Achievement". It was decided, therefore, to search for another model, one which would allow a solution to the guessing issue within more realistic constraints. A first attempt in this search is described in "An Extension of the Dirichlet-Multinomial Model that Allows True Score and Guessing to be Correlated." The new model had theoretical advantages over existing models, but there was no convincing evidence that it had any practical advantages, and after considerable review, it was abandoned.

The next attempt was based on an answer-until-correct scoring model. This solution is described in "Some Empirical and Theoretical Results on an Answer-Until-Correct Scoring Model". All indications are that this new model substantially improves on existing procedures, both theoretically and empirically. However, in a very few instances, some of the items used in the study seemed inconsistent with the assumptions being made. Accordingly, another empirical study was conducted to see whether an additional model would "explain" those remaining items. The results of this study are contained in "Some New Results on an Answer-Until-Correct Scoring Procedure". At the same time, it was also thought desirable to develop a new reliability coefficient that reflects the effectiveness of the distractors being used, as an attempt to integrate the main substantive areas under review. The first step toward this goal is described in "Using k out of n System Reliability to Study and Characterize Tests". However, the reasonableness of certain requisite assumptions was not uniformly stable, and so additional work was undertaken to find a way of improving this situation. A procedure for doing this is shown in "Bounds on the k out of n Reliability of a Test, and an Exact Test for Random Guessing".

In addition, a related concern of reliability is the matter of test length. Two projects previously funded by NIE include approaches to criterion-referenced tests, and determining test length. Our new results have important implications in both these areas, which are described in "Determining the Length of a Criterion-Referenced Test when an Answer-Until-Correct Scoring Procedure is Used", in "A Closed Sequential Procedure for Comparing the Binomial Distribution to a Standard" and in "A Closed Sequential Procedure for Answer-Until-Correct Tests".

In this general area of reliability, the problem of reliable selection also occurs, that is, techniques for identifying the t best of k examinees. An existing procedure is usually impractical because it might require too many items, a test length issue. A step toward solving this problem is to develop retrospective methods, and some results on how this might be done are described in "Approximating the Probability of Identifying the Most Effective Treatment for the Case of Normal Distributions Having Unknown and Unequal Variances." Additional materials generated this year are:

--A Cautionary Note on Estimating the Reliability of a Mastery Test with the Beta Binomial Model

--Methods and Recent Advances in Measuring Achievement; A Response to Molenaar

Each of these papers is provided in the following pages.

The identification of distractors

Our original plan for analyzing distractors is described in "Analyzing the Distractors of Multiple-Choice Test Items or Partitioning Multinomial Cell Probabilities with Respect to a Standard." However, this approach proved to be unsatisfactory on several grounds. In particular, it did not give a direct measure of how effective the distractors really are. One possibility considered for this particular issue was to analyze how distractors behave in the context of the answer-unti-correct test format. Two procedures were proposed and described in "Solving Measurement Problems with an Answer-Until-Correct Scoring Procedure." A problem that remained was determining whether the assumptions made were reasonable. This was empirically investigated in "Some Empirical and Theoretical Results on an Answer-Until-Correct Scoring Procedure", and in "Some New Results on an Answer-Until-Correct Scoring Procedure."

Next, it was deemed important to consider how distractors might be analyzed in terms of their relation to the n items on a test. This work was explicated in "Using k out of n System Reliability to Study and Characterize Tests" and in "Bounds on the k out of n Reliability of a Test. Additional work on distractors is described in "A Polarization Test for Making Inferences About the Entropy of Muultiple-Choice Tests", and in Analyzing the Distractors of Multiple-Choice Test Items or Partitioning Multinomial Cell Probabilities with Respect to a Standard.

METHODS AND RECENT ADVANCES IN
MEASURING ACHIEVEMENT:
A RESPONSE TO MOLENAAR

Rand R. Wilcox

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California . Los Angeles

and the

DEPARTMENT OF PSYCHOLOGY
University of Southern California

ABSTRACT

Commenting on a paper I published in this journal (Wilcox, 1979a), Molenaar (1981) has raised some questions about the usefulness and feasibility of measuring achievement with latent structure models. In the last two or three years, considerable progress has been made regarding the issues mentioned by Molenaar. The purpose of this note is to indicate the progress that has been made, to describe alternative solutions that have been recently proposed, and to comment on some of Molenaar's suggestions on how the model might be improved.

1. INTRODUCTION

Commenting on a paper I published in this journal (Wilcox, 1979a), Molenaar (1981) has raised some important issues related to measuring achievement with latent structure models. The purpose of this note is to briefly outline where we now stand in regard to the concerns expressed by Molenaar. Before doing so, let me establish some notation, and make some opening remarks.

Suppose we have a domain of skills and a single examinee. Let ζ be the proportion of skills the examinee has acquired. Further suppose that every skill is represented by one or more items. Let β=Pr(correct response | the examinee does not know) when the examinee answers an item corresponding to a randomly sampled skill. Finally, let γ be the joint probability of not knowing and being correct. Thus, $\gamma = \beta(1-\zeta)$.

The above model is based on what I call Type II guessing. It is important to realize that the latent structure models referenced in my paper (Wilcox, 1979, p. 62) are based on Type I guessing. That is, guessing is defined in terms of a single skill and a population of examinees. The purpose of the first section of my paper was to show that we can interchange the role of items and examinees to estimate Type II guessing which in turn makes it possible to solve the problems described in sections 2 and 3.

2. PAIRWISE EQUIVALENT ITEMS

The first issue raised by Molenaar is about the notion of equivalent items. Two items that measure the same skill are said to be equivalent if an examinee knows both or neither one. Molenaar points out that equivalent items might exist in some instances, but there are situations where the creation of equivalent items is difficult or even impossible. He is, of course, correct.

Two aspects of this problem need to be addressed. The first is to indicate six ways we can empirically check whether two or more items are equivalent. The second is to briefly comment on four alternative approaches to the problem of guessing.

The first and perhaps most obvious approach to checking whether items are equivalent is to apply the usual chi-square goodness-of-fit test to the latent structure model being used. Macready and Dayton (1977) illustrate this for a model based on Type I guessing and equivalent items. We note that a good fit to their data was obtained.

Observe, though, that a poor fit does not necessarily imply that items are not equivalent. It might mean that a more general model is needed. For example, we might assume that Pr(incorrect response | examinee knows)>0 (Macready and Dayton, 1977).

The second approach is to estimate an index that measures equivalence (Baker and Hubert, 1977).

Another way to check whether items are equivalent is to use latent partition analysis in the manner proposed by Hartke (1978).

The fourth solution is to first assume that one of the items is hierarchically related to the second. A test of this assumption is given by White and Clark (1973). (See, also, Dayton and Macready, 1976.) If the items are indeed equivalent, one of the parameters in the resulting latent structure model, say δ, will be zero (Wilcox, 1980a). If we assume $\beta>0$, a test of the hypothesis that $\delta=0$ can be made by testing the equality of two cell probabilities in a 2x2 contingency table. This can be done with McNemar's test. Some results on the power of McNemar's test are given in Wilcox (1977a).

A fifth check on the assumption of equivalent item pairs can be made if we assume β is bounded above by some constant less than 1. For example, if $\beta \leq \frac{1}{2}$, then from Wilcox (1979a, p. 64) it follows that two cell probabilities in a 2x2 contingency table must be less than or equal to $\frac{1}{4}$. One way to check this assumption is described by Wilcox (in press, a).

Finally, assuming $\beta \leq \frac{1}{2}$ also implies that for a randomly sampled pair of equivalent items, the probability of a correct-incorrect response (and the probability of an incorrect-correct response) is less than or equal to the probability of two incorrect responses. This inequality is easily verified by again referring to Wilcox (1979a, p. 64). Robertson (1978) describes a test of this assumption.

Alternative Approaches to Guessing

Suppose that empirical evidence does not support the assumption of equivalent items or that we decide a priori that equivalent items do not exist. In this case we have four alternative approaches to the problem of guessing. The first is to use completion items. This might eliminate guessing, but errors at the item level might still exist (Harris et al., 1980; Macready and Dayton, 1977). Also, in many situations, scoring completion items is economically infeasible.

If multiple-choice items must be used, the second alternative is to assume guessing is at random. Lord and Novick (1968, p. 309) note that this assumption can seldom be seriously entertained. Empirical investigations on the usual correction-for-guessing formula score support this view (Cross and Frary, 1977; Bliss, 1980). We might assume guessing is at random anyway, but this can have serious consequences in terms of the design and accuracy of a test (Weitzman, 1970; Wilcox, 1980b, 1980c).

Another approach is to assume hierarchically related items are available. The resulting model includes equivalent items as a special case (e.g., Wilcox, 1980a). Dayton and Macready (1976) describe a general framework for handling hierarchically related items. For an even more general model, see Dayton and Macready (1980).

The fourth alternative is to use an answer-until-correct scoring procedure proposed by Wilcox (1981). (For a related scoring rule, see Brown, 1965.) Suppose multiple-choice test items are used with t alternatives from which to choose, one of which is correct. An examinee chooses alternatives until the correct response is identified. Assume the examinee can eliminate i distractors from consideration when the correct response is not known, i=0,1,...,t-2. Following Horst (1933), we also assume that the examinee guesses at random from among those distractors that are not eliminated. For a specific examinee let p_i be the probability of choosing the correct response on the ith attempt of a randomly selected item (i=1,...,t), and let ζ_j (j=0,...,t-2) be the proportion of items in the item pool for which the examinee can eliminate j distractors. The probability of a correct on the first attempt is

$$p_1 = \zeta + \sum_{j=0}^{t-2} \zeta_j/(t-j)$$

and

$$p_i = \sum_{j=0}^{t-i} \zeta_j/(t-j) \qquad (i=2,\ldots,t).$$

The model assumes

$$p_1 \geq p_2 \geq \cdots \geq p_t \qquad (2.1)$$

which can be tested (Robertson, 1978). When (2.1) is assumed, maximum likelihood estimates of the p_i's are easily obtained using the "pool-adjacent-violators" algorithm (Barlow, et al., 1972, pp. 13-18). If $\hat{p}_1$ and $\hat{p}_2$ are the usual sample mean estimates of p_1 and p_2, it follows that a maximum likelihood estimate of ζ is $\hat{\zeta}=\hat{p}_1-\hat{p}_2$ if $\hat{p}_1 \geq \hat{p}_2$, and if $\hat{p}_1 < \hat{p}_2$, the estimate is zero (Zehna, 1966).

In addition to correcting for partial information, the model can solve several other measurement problems (Wilcox, 1981b, 1980e). Suppose, for example, we have an n-item test, and that ξ is the expected number of items for which we correctly determine whether the typical examinee knows the correct response. Using results in the engineering literature on "system reliability" it is possible to make inferences about whether ξ is large or small (Wilcox, 1980e).

Before concluding this section we make the important observation that latent structure models based on the notion of equivalent items have been successfully applied to real data sets (Macready & Dayton, 1977; Harris & Pearlman, 1978). More recently, Professor C. W. Harris and his collegues made extensive use of these models to measure the arithmetic achievement of students in various grade levels. Examinees were tested every week over a period of many weeks. All indications are that the models are indeed useful.

Finally, Molenaar writes that the estimates of ζ and β are unbiased only if the selected pairs of equivalent items are representative of the item pool. Actually, the estimates appear to be always biased whether we have a random sample or not. However, we do get maximum likelihood estimates as long as the estimates have an admissible value (Wilcox, 1977).

3. THE MULTINOMIAL MODEL

In the next section of Molenaar's paper, he turns his attention to the multinomial model. Suppose an examinee responds to n items, none of which are equivalent. (A strong true score model for equivalent item pairs is described in Wilcox, 1981.) Still considering only a single examinee, let y be the number of items he/she knows, and let z be the number of items not known but guessed correctly. My paper (Wilcox, 1979a) considers a bivariate analog of the binomial error model (Keats and Lord, 1962; Lord, 1965; Lord and Novick, 1968, chapter 23). In particular, I assume that the joint density of y and z is

$$f(y,z \mid \zeta,\gamma) = \frac{n!\zeta^y\gamma^z(1-\zeta-\gamma)^{n-y-z}}{y!z!(n-y-z)!}, \tag{3.1}$$

where n is the number of items on the test. Ordinarily we cannot make inferences about ζ and γ, but as already indicated, we can make inferences about them when equivalent or hierarchically related items are available, or when an answer-until-correct scoring procedure is used.

Of course we can assume guessing is random (see in particular, Morrison and Brockway, 1979), but I have already described the problems with this. However, we can empirically test whether guessing is at random (Weitzman, 1970; Wilcox, 1981b). When an answer-until-correct scoring procedure is used, this corresponds to testing whether $p_2=p_3= \dots =p_t$. If guessing is not at random, perhaps infrequently chosen distractors could be modified or replaced so that this assumption is more realistic. In this case, results in Morrison and Brockway (1979), and Molenaar (1977), might be applied. Wilcox (in press a) gives some results that might be useful in identifying those distractors that are infrequently chosen. Note that we can also measure how far away guessing is from being random (Wilcox, 1981b), and we can empirically determine how many distractors are needed when testing a particular population of examinees (Wilcox, 1980e).

4. THE DIRICHLET PRIOR

In Wilcox (1979a), I assume that ζ and γ have a bivariate Dirichlet density given by

$$g(\zeta,\gamma) = \frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1)\Gamma(\nu_2)\Gamma(\nu_3)}\ \zeta^{\nu_1-1}\ \gamma^{\nu_2-1}\ (1-\zeta-\gamma)^{\nu_3-1} \tag{4.1}$$

If we can estimate ζ and γ for N randomly sampled examinees, we can estimate the ν_i's. I used equivalent items in Wilcox (1979a) to do this, but as already noted, two other approaches are now available which do not assume random guessing.

Let $\hat{\zeta}_i$ and $\hat{\beta}_i$ be the maximum likelihood estimates of ζ and β, respectively, for the ith randomly sampled examinee $(i=1,\ldots,N)$. Molenaar raises the interesting question of whether we can improve upon $\hat{\zeta}_i$ and $\hat{\beta}_i$ by shrinking their values toward each other. Molenaar alludes to the possibility of using Kelley's regression estimate of true score. If "better" estimates of ζ and β are available, we might be able to get improved estimates of the hyperparameters ν_1, ν_2 and ν_3. If an ensemble squared error loss function is believed to be appropriate when estimating ζ(and β), there is reason to hope that such a procedure might improve upon the maximum likelihood estimate of ζ (and β) used in my paper (e.g., Efron and Morris, 1973; Wilcox, 1978a). Griffin and Krutchkoff (1971) show that Kelley's regression estimate of a parameter is the optimal linear estimate under squared error loss if we start with an unbiased estimate of the parameter. However, the estimates of ζ (and β) is biased, and so it is not clear whether Kelley's regression equation will help improve my estimates of ζ and β. We might use Kelley's regression estimate anyway, but the efficacy of this needs to be checked. For an alternative way of possibly improving the estimation of ζ (and β), see Wilcox (1980c).

Molenaar also implies that using Kelley's regression estimate of ζ and β might also improve the estimates of the ν_i's. There is, unfortunately, no evidence that this is ever the case. In an unpublished report, I tried a similar tactic in a situation where unbiased estimates of a parameter were available, but the results were not overly convincing.

Next Molenaar comments on the numerical example in my paper. To estimate the ν_i's, I used artificially generated data on 1,000 examinees taking 100 pairs of equivalent items. Molenaar inferred that a large number of items and examinees are needed to get reasonably accurate estimates. It should be pointed out, however, that the number of examinees and items was completely arbitrary. Just how accurate an estimate of the ν_i's we get with a smaller number of items is unknown. We would, of course, expect the accuracy of the estimates to depend on actual values of ζ and β (cf. Wilcox, 1980a). From Wilcox (1979b) we would also expect to find instances where a moderate number of examinees would give wildly inaccurate results. Such situations might be rare, but this has not been established. The main point is that currently there is no information on how many items should be used when applying the model.

Note that for reasons given by Mosimann (1962) a slight modification of the estimates of the ν_i's used in Wilcox (1979a) might be desirable. Mosimann (1962) describes the procedure, and Wilcox (1981a) indicates how to apply it to the case where we have pairs of equivalent items. Any future investigations on estimating the ν_i's should include this procedure.

Molenaar also raises the important issue that the binomial error model (and consequently the multinomial model) implies that all items have the same level of difficulty. From a theoretical point of view, this restriction of the model is unacceptable. A simple way to eliminate this problem is to use an approximation to the compound binomial distribution (Lord, 1965). However, for many purposes, this seems to be unnecessary (Lord, 1965; Algina and Noe, 1978; Wilcox, 1977b, 1978a). Also the beta-binomial model has given good results in other empirical investigations (Gross and Shulman, 1980; Subkoviak, 1978a). Since the beta-binomial model appears to be both useful and robust in certain respects, there is hope that the Dirichlet-multinomial will share the same properties since it is the multivariate-analog of the beta-binomial model. Some evidence for this is given in Wilcox (in press b) where the Dirichlet-multinomial model was applied to real data, but more work needs to be done. For further discussions of the binomial and beta-binomial models, see Wilcox (1981).

Molenaar suggests that to accomodate unequal item difficulties, we might use the Rasch model. (For a review of latent trait models, see Hambleton et al., 1978; and for a review of some recent developments on the Rasch model, see Wainer et al., 1980.) However, this model does not yield an estimate of ζ, at least not in any way that has been demonstrated, and it ignores the problem of guessing. Some latent trait models--but not the Rasch model--have what is sometimes called a guessing parameter. This is just the lower asymptote of the item characteristic curve. Note, however, that this is different from the notion of Type I and Type II guessing. Thus, the Rasch model is unable to solve any of the measurement problems described in Wilcox (1981b, 1980e). No claim is being made that latent trait models are useless, nor do I believe that latent trait and latent structure models are in competition with one another--the point is that they answer different questions. For further critical remarks regarding the Rasch model, see Lord (1974).

5. MODEL ADEQUACY

Molenaar objects to the implication of the Dirichlet-multinomial model that ζ and β are independent over the population of examinees, and from a theoretical point of view, he is, of course, correct. As Molenaar puts it, "One wonders whether a person who knows many items from the domain will also be more clever in guessing the remaining ones if only by the 'warm glow of success'"? The first point is that if we throw out the model because ζ and β are independent, we must throw out the random guessing model as well since ζ and β are again independent. The second point is that when addressing a particular measurement problem the seriousness of assuming ζ and β to be independent is not known.

To allow ζ and β to be correlated, there appears to be three possibilities. The first is to replace (3.1) with

$$g(\zeta,\gamma) = \sum_{j=0}^{\infty} \frac{c_j\tau^j}{\Psi(\tau)} \cdot \frac{\Gamma(\nu_1+\nu_2+\nu_3+j)}{\Gamma(\nu_1+j)\Gamma(\nu_2)\Gamma(\nu_3)} \zeta^{\nu_1+j-1}\, \gamma^{\nu_2-1}(1-\zeta-\gamma)^{\nu_3-1} \quad (4.1)$$

where c_j is a constant depending on j, but not τ, τ is an unknown parameter, and Ψ is a function of τ (Wilcox, 1981a). The density (4.1) contains (3.1) as a special case. Moreover, if ζ and β are assumed to be continuous, they are independent if and only if (4.1) reduces to (3.1). One choice for c_j and Ψ is $c_j=(j!)^{-1}$ and $\Psi(\tau)=e^{\tau}$, in which case the marginal density of ζ belongs to the non-central beta family. Assuming (4.1) holds, let $r=\nu_1$, $s=\nu_2+\nu_3$, and let E_y mean expectation with respect to the probability function

$$f(y) = \frac{c_y\tau^y}{\Psi(\tau)} \qquad (y=0,1,\ldots).$$

The first four moments of the marginal density of ζ are

$$\mu_1 = 1-sE_y\left(\frac{1}{r+s+y}\right)$$

$$\mu_2 = s-(s-1)\mu_1-(s+1)sE_y\left(\frac{1}{r+s+1+y}\right)$$

$$\mu_3 = \tfrac{1}{2}\{2-s(s-1)(s-2)E_y\left(\frac{1}{r+s+y}\right) + 2s(s-1)(s+1)E_y\left(\frac{1}{r+s+1+y}\right)$$

$$- s(s+1)(s+2)\, E_y\left(\frac{1}{r+s+2+y}\right)\}$$

$$\mu_4 = \mu_3 - \frac{s}{3}\{\mu_3-d_1\} \qquad (4.2)$$

where

$$d_1 = \tfrac{1}{2}\{r+E(y)-2s-s(s-1)(s+1)Ey\left[\frac{1}{r+s+1+y}\right]$$

$$+ s(s+1)(s+2)\; E_y\left[\frac{1}{r+s+2+y}\right] \; - d_2\},$$

$$E(y) = \tau\Psi'(\tau)/\Psi(\tau), \text{ and}$$

$$d_2 = r + E(y)-2(s+1)+(s+1)(s+2)\; E\left[\frac{1}{r+s+y+2}\right]$$

$$+ (s+1)(s+3)\; E\left[\frac{1}{r+s+y+3}\right]$$

$$+ (s+1)^2\; \{E\left[\frac{1}{r+s+y+3}\right] \; - (s+2)\; E\left[\frac{1}{r+s+2+y} - \frac{1}{r+s+3+y}\right]\}$$

Note that there is no need to evaluate $E(y)$ when calculating μ_4 since $E(y)$ cancels out.

It can be seen that $E_y(t^y)=\Psi(\tau t)/\Psi(\tau)$ and so

$$E_y\left[\frac{1}{r+s+k-1+y}\right] = \int_0^1 u^{r+s+k-1}\; \Psi(\tau u)du$$

for any integer $k \geq 0$ (Chao and Strawderman, 1972). The integral in this last expression can be evaluated with IMSL (1975) subroutine DECADRE. Thus, the method of moments might be used to estimate the parameters in (4.1). It should be stressed, however, that the practical advantages of using (4.1) are not known.

The second approach to allowing ζ and β to be correlated is to follow the suggestion of Aitchison and Shen (1980) and replace (3.1) with a logistic normal distribution. However, the moments are not reducible to any simple form which makes this approach impractical for the problem at hand. For alternative generalizations of (3.1), see the papers cited in Wilcox (1979a).

The third approach is to apply Dirichlet-multinomial to an answer-until-correct scoring procedure. This, and other models, is now being tried out on some real data. The results should be available in the near future.

6. CONCLUDING REMARKS

The goal in Wilcox (1979a) was to suggest a strong true-score model that allows guessing to vary over a population of examinees. Another motivation for the model was that there are real situations where equivalent items are assumed (e.g., Wilcox, in press b), but previously there was no strong true-score model for handling this case.

Molenaar has raised some important concerns about whether the problem of guessing has been satisfactorily dealt with. Considerable progress has been made since my paper was published, but I still agree with him that more work needs to be done. The important point of this paper is that today we have several methods for dealing with guessing without assuming it is at random. Moreover, each solution can be empirically checked in several ways. Early attempts at correcting for guessing were based on rather restrictive assumptions, but there seems to be situations where these assumptions are appropriate. More recent solutions are based on weaker assumptions, but we need more experience with them before they are routinely applied. As previously indicated, an empirical investigation of an answer-until-correct scoring procedure is currently underway which should partially correct this problem.

REFERENCES

Aitchison, J., & Shen, S. M. (1978). Logistic-normal distributions: Some properties and uses. Biometrika, 67, 261-272.

Algina, J., & Noe, M. J. (1978). A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 15, 101-110.

Baker, F. B., & Hubert, L. J. (1977). Inference procedures for ordering theory. Journal of Educational Statistics, 2, 217-233.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. (1972). Statistical inference under order restrictions. New York: Wiley.

Bliss, L. B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Brown, J. (1965). Multiple response evaluation of discrimination. The British Journal of Mathematical and Statistical Psychology, 18, 125-137.

Chao, M. T., & Strawderman, W. E. (1972). Negative moments of positive random variables. Journal of the American Statistical Association, 67, 429-431.

Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 14, 313-321.

Dayton, C. M., & Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. Psychometrika, 41, 189-204.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors. Journal of the American Statistical Association, 68, 117-130.

Griffin, B. S., & Krutchkoff, R. G. (1971). Optimal linear estimators: an empirical Bayes version with application to the binomial distribution. Biometrika, 58, 195-201.

Gross, A. L., & Shulman, V. (1980). The applicability of the beta-binomial model for criterion-referenced testing. Journal of Educational Measurement, 17, 195-202.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 48, 467-510.

Hartke, A. R. (1978). The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 1978, 15, 43-47.

Harris, C. S., & Pearlman, A. (1978). An index for a domain of completion or short answer items. Journal of Educational Statistics, 3, 285-304.

Horst, P. (1933). The difficulty of a multiple-choice test item. Journal of Educational Psychology, 24, 229-232.

IMSL Library 1. (1975). Volume II. Houston: International Mathematical and Statistical Libraries.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika, 27, 59-72.

Lord, F. M. (1965). A strong true-score theory, with applications. Psychometrika, 30, 239-270.

Lord, F. M. (1974). An individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes (Eds.) Contemporary developments in mathematical psychology, Volume II. San Francisco: Freeman.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 2, 99-120.

Molenaar, I. (1977). On Bayesian formula scores for random guessing in multiple choice tests. British Journal of Mathematical and Statistical Psychology, 30, 79-89.

Molenaar, I. (1981). On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology, 34.

Morrison, D. B., & Brockway, G. A modified beta-binomial model with applications to multiple choice and taste tests. Psychometrika, 44, 427-442.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika, 49, 65-82.

Robertson, T. (1978). Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 73, 197-202.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 15, 111-116.

Wainer, H., Morgan, A., & Gustafsson, J. (1980). A review of estimation procedures for the Rasch model with an edge toward longish tests. Journal of Educational Statistics, 5, 35-64.

Weitzman, R. A. (1970). Ideal multiple-choice items. Journal of the American Statistical Association, 65, 71-89.

White, R. T., & Clark, R. M. (1973). A test of inclusion which allows for errors of measurement. Psychometrika, 38, 77-86.

Wilcox, R. R. (1977a). New methods for studying stability. In C. W. Harris, A. Pearlman, & R. Wilcox, Achievement Tests Items--Methods of Study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California.

Wilcox, R. R. (1977b). Estimating the likelihood of a false-positive and false-negative decision with a mastery test: An empirical Bayes approach. Journal of Educational Statistics, 2, 289-307.

Wilcox, R. R. (1978). Estimating true score in the compound binomial error model. Psychometrika, 43, 245-258.

Wilcox, R. R. (1979a). Achievement tests and latent structure models. British Journal of Mathematical and Statistical Psychology, 32, 61-71.

Wilcox, R. R. (1979b). Estimating the parameters of the beta-binomial distribution. Educational and Psychological Measurement, 31, 527-535.

Wilcox, R. R. (1980a). Some results and comments on using latent structure models to measure achievement. Educational and Psychological Measurement, 40, 645-658.

Wilcox, R. R. (1980b). An approach to measuring the achievement or proficiency of an examinee. Applied Psychological Measurement, 4, 241-251.

Wilcox, R. R. (1980_c_). Determining the length of a criterion-referenced test. _Applied Psychological Measurement_, to appear.

Wilcox, R. R. (1980_d_). Solving measurement problems with an answer-until-correct scoring procedure. Center for the Study of Evaluation, University of California, Los Angeles.

Wilcox, R. R. (1980_e_). Using results on k out of n system reliability to study and characterize tests. Center for the Study of Evaluation. University of California, Los Angeles.

Wilcox, R. R. (1981). A review of the beta-binomial model and its extensions. _Journal of Educational Statistics_, to appear.

Wilcox, R. R. (in press, _a_). Analyzing the distractors of multiple-choice test items or partitioning multinomial cell probabilities with respect to a standard. _Educational and Psychological Measurement_.

Wilcox, R. R. (in press, _b_). The single administration estimate of the proportion of agreement of a proficiency test scored with a latent structure model. _Educational and Psychological Measurement_.

Zehna, P. W. (1966). Invariance of maximum likelihood estimation. _Annals of Mathematical Statistics_, _37_, 744.

An Extension of the Dirichlet-Multinomial Model that Allows True Score and Guessing to be Correlated

Rand R. Wilcox

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California . Los Angeles

Abstract

Most strong true-score models assume that when an examinee does not know the correct response to a test item, the probability of guessing, say β, is independent of an examinee's true score. In fact, it is common practice to make the more restrictive assumption that β is the same known constant for every examinee. One exception is the Dirichlet-multinomial model; but true score and guessing are still assumed to be independent. This paper describes an extension of the Dirichlet-multinomial model that allows true score and guessing to be correlated.

Consider a multiple-choice test item designed to determine whether an examinee has acquired a particular skill. An obvious problem is that an examinee can give a correct response without knowing the answer; yet, in many situations, it is economically infeasible to use completion items in an attempt to correct this difficulty. On the otherhand, guessing can have serious implications for certain types of achievement tests (e.g., Wilcox, 1980a, 1980b). Thus, it is natural to search for scoring procedures and probability models that take guessing into account.

Suppose a multiple-choice test item has t alternatives consisting of t-1 distractors and one correct response. Typically, the problem of guessing is handled by assuming that β=Pr (correct response | examinee does not know) = 1/t, i.e., guessing is at random (e.g., Hamilton, 1950; Chernoff, 1962; Duncan, 1974; Morrison and Brockway, 1979). There are at least two serious objections to this approach. First, it is unrealistic to assume that every examinee has the same probability of guessing. For example, some examinees might be able to eliminate one or more distractors from consideration without knowing the correct response. In this case we would expect to have $\beta>1/t$. We might assume $\beta\approx 1/t$ but in some instances this does not yield satisfactory results (Wilcox, 1980b). The second objection to setting $\beta=1/t$ is the implication that true score and guessing are independent. As argued by Frary (1969), we would expect this assumption to be false. Of course, if we set $\beta=0$, we still have this problem.

Let ζ be the proportion of skills among a domain of skills that an examinee has acquired and set $\gamma=(1-\zeta)\beta$. Wilcox (1979) proposed a solution to the first problem by assuming that, over the population of examinees

ζ and γ have a bivariate Dirichlet distribution given by

$$(1.0) \qquad \frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1)\Gamma(\nu_2)\Gamma(\nu_3)} \zeta^{\nu_1-1} \gamma^{\nu_2-1} (1-\zeta-\gamma)^{\nu_3-1}$$

where $\nu_i>0$, $i=1,2,3$ are unknown parameters and Γ is the gamma function. It was also assumed that the probability of x correct responses for an examinee taking an n-item test is

$$(1.1) \qquad \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

where $\theta=\zeta+\gamma$ is the examinee's percent correct true score. However, the second problem remains since (1.0) implies that ζ and β are independent. The restrictive nature of (1.0) has also concerned statisticians (e.g., James, 1975; Connor and Mosimann, 1969; and Antelman, 1972) but the proposed generalizations of the Dirichlet distribution have proven to be less than satisfactory. The purpose of this paper is to describe a broad class of distributions that contains (1.0) as a special case, and which allow ζ and β to be correlated. Our general results are illustrated for the special case where the marginal distribution of ζ is non-central beta. Before continuing, however, it is convenient to examine various extensions of

$$(1.2) \qquad \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \theta^{r-1} (1-\theta)^{j-1},$$

the beta distribution.

2. A Generalization of the Beta Distribution

In this section we describe a family of probability density functions where (1.2) is "mixed" by a distribution that belongs to a large class of discrete probability functions. We then indicate how our results might be applied when a particular generalization of the beta density is used to approximate $g(\theta)$, the distribution of θ over the population of examinees.

Consider a random variable Y having the probability function.

$$(2.1) \qquad P(Y=y) = \frac{c_y \tau^y}{\Psi(\tau)} , \; y=0,1,\ldots.$$

where τ is an unknown parameter, c_y is a constant depending on y but not τ, and Ψ is a function of τ. Expression (2.1) is referred to as a power series distribution by Noack (1950). There are even more general discrete distributions that contain the power series distribution as a special case, (e.g., Patil, 1962; Gupta, 1974), but they are not discussed here since (2.1) is of sufficient generality for present purposes.

Consider

$$(2.2) \qquad g(\theta) = \sum_{j=0}^{\infty} \frac{c_j \tau^j}{\Psi(\tau)} \; \frac{\Gamma(r+s+j)}{\Gamma(r+j)\Gamma(s)} \; \theta^{r+j-1} \; (1-\theta)^{s-1} \; .$$

It is readily verified that (2.2) has the properties of a probability density function, i.e., it is non-negative and it integrates to one. If, for example, we set $c_j=(j!)^{-1}$ and $\Psi(\tau)=e^{-\tau}$ we get the non-central beta distribution (see, e.g., Seber, 1963), and, if in addition $\tau=0$, (2.2) reduces to (1.2). The first three moments of (2.3) are, respectively,

$$(2.3) \qquad \mu_1 = 1-sE_y\left[\frac{1}{r+s+y}\right]$$

$$(2.4) \qquad \mu_2 = s-(s-1)\mu_1-(s+1)s \; E_y\left[\frac{1}{r+s+1+y}\right]$$

$$(2.5) \qquad \mu_3 = \tfrac{1}{2}\Big[2-s(s-1)(s-2) \; E_y\left[\frac{1}{r+s+y}\right] + 2s(s-1)(s+1)E_y \left[\frac{1}{r+s+1+y}\right]$$

$$- \; s(s+1)(s+2) \; E_y \left[\frac{1}{r+s+2+y}\right]\Big]$$

where E_y denotes expectation with respect to the density given in expression (2.1). It can be seen that $E_y(t^y)=\Psi(\tau t)/\Psi(\tau)$ and so from Chao and Strawderman (1972) it follows that

$$(2.6) \qquad E_y\left(\frac{1}{r+s+k-1+y}\right) = \int_0^1 u^{r+s+k-1}\Psi(\tau u)du.$$

for any integer $k \geq 0$. For a detailed derivation of these moments, see Wilcox (1980c).

Again omitting the tedious algebra, it can also be shown that

$$(2.7) \qquad \mu_4 = \mu_3 - \frac{s}{3}\ [\mu_3 - d_1]$$

where

$$\begin{aligned} d_1 = \tfrac{1}{2}\,[&r+E(y)-2s-s(s-1)(s+1)\ E\left(\frac{1}{r+s+1+y}\right) \\ &+ s(s+1)(s+2)\ E_y\left(\frac{1}{r+s+2+y}\right) - d_2], \end{aligned}$$

$$E(y) = \tau\Psi'(\tau)/\Psi(\tau), \text{ and}$$

$$\begin{aligned} d_2 = r + E(y) - 2(s+1) &+ (s+1)(s+2)\ E\left(\frac{1}{r+s+y+2}\right) \\ &+ (s+1)(s+3)\ E\left(\frac{1}{r+s+y+3}\right) \\ &+ (s+1)^2\ [E\left(\frac{1}{r+s+y+3}\right) - (s+2)\ E\left(\frac{1}{r+s+2+y} - \frac{1}{r+s+3+y}\right)] \end{aligned}$$

Note that there is no need to evaluate $E(y)$ when calculating μ_4 since $E(y)$ cancels out.

Some special cases. To illustrate the results given above, suppose

$$P(Y=y) = \frac{\Gamma(\delta+y)}{y!\ \Gamma(\delta)}\ \tau^{y}(1-\tau)\ ,\ y=0,1,\ldots. \tag{2.8}$$

where $\delta>0$ and $0>\tau>1$. In terms of (2.1) we get this distribution by setting $c_y=\Gamma(\delta+y)/(y!\Gamma(\delta))$ and $\Psi(\tau)=(1-\tau)^{-\delta}$. Thus, (2.5) becomes

$$E_y\left(\frac{1}{r+s+k-1+y}\right) = \int_0^1 u^{r+s+k-1}\ [(1-\tau)/(1-u\tau)]\ du.$$

Hence, from expressions (2.3) - (2.5) and (2.7) we have the first four moments of $g(\theta)$.

As another illustration, suppose we replace (2.8) with the hyper-Poisson probability function (Bardwell & Crow, 1964) given by

$$P(Y=y) = \frac{\Gamma(\delta)\ \tau^{y}}{{}_1F_1(\delta,\tau)\ \Gamma(\delta+y)}\ ,\ y=0,1,\ldots.$$

where $\delta>0$, $\tau>0$ and ${}_1F_1(\delta,\tau) = 1 + \frac{\tau}{\delta} + \frac{\tau^2}{\delta(1+\delta)} + \cdots$ is a special case of the confluent hypergeometric series. In this instance

$$g(\theta) = \sum_{j=0}^{\infty} \frac{\Gamma(\delta)\ \tau^{j}\ \Gamma(r+s+j)\ \theta^{r+j+1}(1-\theta)^{s-1}}{{}_1F_1(\delta,\tau)\ \Gamma(\delta+j)\ \Gamma(r+j)\ \Gamma(s)}$$

Setting $\Psi(\tau)={}_1F_1(\delta,\tau)$ for fixed δ, the value of $E\left(\frac{1}{r+s+k-y}\right)$ is given by (2.6). Again we can determine the first four moments of $g(\theta)$ with (2.3)-(2.5) and (2.7).

3. The Non-Central Beta Distribution

Before describing a model that allows ζ and β to be correlated, it is helpful to consider how the results of section 2 can be used to estimate $g(\theta)$. We do this for the case where θ is assumed to have a non-central beta distribution given by

$$(3.1) \qquad g(\theta) = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!} \frac{\Gamma(r+j+s)}{\Gamma(r+j)\Gamma(s)} \theta^{r+j-1} (1-\theta)^{s-1}$$

where $\lambda>0$, $r>0$ and $s>0$ are parameters to be determined. As previously noted, (3.1) is a special case of (2.2). The corresponding marginal distribution of observed scores, assuming (1.1) holds, is

$$h(x) = \sum_{j=0}^{\infty} \frac{B(r+j+x,\ n+s-x)}{(n+1)\ B(r+j,\ s)\ B(x+1,\ n+1-x)}$$

where $B(r, s) = [\Gamma(r)\ \Gamma(s)]/\Gamma(r+s)$. We also note that if y is the observed score on a randomly parallel test having n_1 items, the joint distribution of x and y is

$$h(x, y) = \int_0^1 \binom{n}{x} \theta^{x} (1-\theta)^{n-x} \binom{n_1}{x} \theta^{y} (1-\theta)^{n_1-y}\ g(\theta)\ d\theta$$

$$= \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!} \frac{B(r+j+x+y,\ n+n_1+s-x-y)}{(n+1)(n_1+1)\ B(r+j,s)\ B(1+x,\ n+1-x)\ B(1+y,\ n_1+1-y)}$$

This last result might be useful in the single administration estimate of a mastery test. (See Huynh, 1976.)

We will need a method of estimating the parameters λ, r and s using the observed scores of a random sample of N examinees. The first step in solving this problem is deriving expressions for the first three moments of the non-central beta distribution. From the previous section we have that

(3.2) $$\mu_1 = 1-se^{-\lambda} \int_0^1 t^{r+s-1} e^{\lambda t}\, dt$$

From Wishart (1932, p. 445) we see that

$$\int_0^1 t^{r+s-1} e^{\lambda t}\, dt$$

$$= \frac{1}{r+s} F(r+s, r+s+1, \lambda)$$

$$= \frac{e^{\lambda}}{r+s} F(1, r+s+1, -\lambda).$$

Where F is the confluent hypergeometric series given by

$$F(a, b, c) = 1 + \frac{a}{1!b} c + \frac{a(a+1)}{2!b(b+1)} c^2 + \ldots$$

Hence, we have that

$$\mu_1 = 1 - \frac{s}{r+s} F(1, r+s+1, -\lambda).$$

Tables and computational procedures described by Abramowitz and Stegun (1972, Chapter 13) can be used to evaluate F which in turn gives us the value μ_1 or the value of μ_1 can be determined by evaluating the integral in (2.3) with IMSL (1975) subroutine DECADRE. Note that for $\lambda=0$ (the beta distribution) expression (3.2) reduces to $r/(r+s)$ as it should.

The second moment about the origin is

(3.3) $$\mu_2 = s-(s-1)\mu_1-(s+1)s\, e^{-\lambda} \int_0^1 t^{r+s} e^{\lambda t}\, dt.$$

Finally, the third moment is

(3.4) $$\tfrac{1}{2} \Big[2-s(s-1)(s-2)\, E\left(\frac{1}{r+s+j}\right) + 2s(s-1)(s+1)\, E\left(\frac{1}{r+s+1+j}\right) - s(s+1)(s+2)\, E\left(\frac{1}{r+s+2+j}\right) \Big]$$

where the expectations are taken with respect to a random variable j having a Poisson distribution with parameter λ. Again referring to Chao and Strawderman (1972)

$$E\,\frac{1}{r+s+k+j} = e^{-\lambda} \int_0^1 t^{r+s+k-1}\, e^{\lambda t}\, dt$$

$$= (r+s+k)^{-1}\, F(1,\ r+s+k+1,\ -\lambda),\ k=0,1,2.$$

As before the integral in this last expression can be evaluated with IMSL subroutine DECADRE.

It is known (e.g., Lord and Novick, 1968, p. 521) that μ_k, the kth moment about the origin of the true score distribution, is equal to

(2.6) $M_k/n^{[k]}$, $k=1, 2, \ldots, n$

where

$$M_k = \sum_{x=0}^{n} x^{[k]}\, h(x)$$

is the kth factorial moment of the marginal distribution of observed scores, and

$$x^{[k]}=x(x-1)\ \ldots\ (x-k+1).$$

Thus, we can use the observed scores of a random sample of N examinees to estimate μ_k with say $\hat{\mu}_k$. Substituting $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\mu}_3$ for μ_1, μ_2, μ_3 respectively in equations (2.3), (2.4) and (2.5) and solving for r, s and λ yields estimates of these parameters say $\hat{r}$, $\hat{s}$ and $\hat{\lambda}$.

At present, the solution to these equations is being obtained using numerical analysis techniques. In particular, we used subroutine ZSYSTM to solve μ_1 and μ_2 for r and s using a fixed value of λ. As initial estimates of r and s we set $\lambda=0$ in which case explicit estimates of r and s are availabe is indicated in the numerical illustration below. With the initial estimates of r, s and λ we computed the corresponding value of μ_3. If this value is not in close agreement with $\hat{\mu}_2$, we increased λ by one, solved for r and s, and again computed the implied value of μ_3. We repeated this process until values of λ, r and s were found that give a good approximation to μ_3.

Numerical illustration. Suppose we have a 5-item test and that f_x examinees received an observed score of x, the values of which are summarized in Table 1.

Table 1

Observed Frequencies on a 5-Item Test

x:	0	1	2	3	4	5
f_x:	23	19	33	15	6	4

The first three moments of the true score distribution were estimated to be .652, .458 and .339 respectively.

Setting λ=0 and using the method of moments, we estimate r and s with

$$\hat{r} = \frac{(\hat{\mu}_1)^2 (1-\hat{\mu}_1)}{\hat{\mu}_2 - \hat{\mu}_1^2} - \mu_1$$

$$\hat{s} = \frac{\hat{\mu}_1 (1-\hat{\mu}_1)^2}{\hat{\mu}_2 - \hat{\mu}_1^2} + \hat{\mu}_1 - 1$$

(e.g., Huynh, 1976; Wilcox, 1977) yielding $\hat{r}$=3.93 and $\hat{s}$=2.04. From expression (2.5), or from standard results on the beta distribution, these values of r, s and λ imply that μ_3=.346, butas previously noted, the estimate, of μ_3 was .339. Therefore, we increased λ to 1 and solved (3.2) and (3.3) for r and s with IMSL (1975) subroutine ZSYSTM yielding $\hat{r}$=3.2876 and $\hat{s}=2.2149$. From (3.4) it follows that $\mu_3=.33896$. Thus, these values of r, s and λ are in reasonably good agreement with the estimated values of μ_1, μ_2 and μ_3. If (3.4) had yielded a number for μ_3 greater than .339, we would have increased λ from 1 to 2 and repeated the process.

4. Extensions to the Dirichlet-Multinomial Model

In this final section, we use the results of the previous two sections to extend the Dirichlet-multinomial model so as to allow ζ and β to be correlated. First, a brief review of this model is in order.

Consider a single examinee responding to n dichotomously scored items randomly sampled from some item pool. Let x be the examinee's number correct score, y be the number of items the examinee knows and z be the number of items that the examinee does not know but guesses the correct response. Let ζ be the proportion of items in the item domain that an examinee knows and let β be the probability of guessing the correct response given that the examinee does not know. It follows that y and z have a multinomial probability function given by

$$\frac{n!\ \zeta^y\ \gamma^z\ (1-\zeta-\gamma)^{n-z-y}}{y!\ z!\ (n-y-z)}$$

where $\gamma=(1-\zeta)\beta$. As previously mentioned, Wilcox (1979) assumes that ζ and γ have a bivariate Dirichlet distribution given by (1.0). The model contains the beta-binomial model as special case (when $\beta=0$) and so in terms of applications, it has all of the appealing features of the beta-binomial model that are described by Lord (1965). An added advantage is that the model allows guessing to vary over the population of examinees. In some cases latent structure models can be used to estimate ζ and β for a specific examinee which in turn makes it possible to apply it to real data. (See Wilcox, 1979, for further details.)

The form of the non-central beta distribution suggests a generalization of (1.0) More specifically we consider replacing (1.0) with

$$(4.1) \qquad g(\zeta, \gamma) = \sum_{j=0}^{\infty} (e^{-\lambda} \lambda^{j}/j!) \frac{\Gamma(\nu_1+\nu_2+\nu_3+j)}{\Gamma(\nu_1+j)\, \Gamma(\nu_2)\, \Gamma(\nu_3)} \zeta^{\nu_1+j-1} \gamma^{\nu_2-1} (1-\zeta-\gamma)^{\nu_3-}$$

It is readily verified that (4.1) is a probability density function. Note that if $\lambda=0$, (4.1) reduces to the Dirichlet distribution and so we expect it to give as good or better an approximation to the joint density of ζ and γ.

From known results about (1.0) it follows that the marginal densities of ζ and γ are non-central beta distributions given by

$$(4.2) \qquad g_1(\zeta) = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!} \frac{\Gamma(\nu_1+\nu_2+\nu_3+j)}{\Gamma(\nu_1+j)\, \Gamma(\nu_2+\nu_3)} \zeta^{\nu_1+j-1} (1-\zeta)^{\nu_2+\nu_3-1} .$$

and

$$(4.3) \qquad g_2(\gamma) = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!} \frac{\Gamma(\nu_1+\nu_2+\nu_3+j)}{\Gamma(\nu_2)\, \Gamma(\nu_1+\nu_3+j)} \gamma^{\nu_2-1} (1-\gamma)^{\nu_1+\nu_3+j-1} .$$

From results given by Ishii and Hayakawa (1960) it can be deduced that the marginal distribution of y and z is

$$(4.4) \qquad p(y, z) = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!} \frac{B(\nu_1+y+j,\ \nu_2+z,\ n+\nu_3-y-z)}{(n+2)(n+1)\, B(\nu_1+j,\nu_2,\nu_3)\, B(1+y,1+z,n+1-y-z)}$$

where $B(a, b, c) = [\Gamma(a)\, \Gamma(b)\, \Gamma(c)]/\Gamma(a+b+c)$.

The density of $x=y+z$ is

$$f(x) = \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!} \frac{B(\nu_1+\nu_2+j+x,\ n+\nu_3-x)}{(n+1)\, B(\nu_1+\nu_2+j,\ \nu_3)\, B(1+x,\ n+1-x)}$$

and the joint distribution of x and ζ is

$$(4.5) \qquad p(x, \zeta) = \binom{n}{x} \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!\, B(\nu_1+j,\nu_2,\nu_3)}$$

$$\left[\sum_{w=0}^{\infty} \binom{x}{w} B(w+\nu_2,\ n-X+\nu_3)\, \zeta^{x-w+\nu_1+j-1} (1-\zeta)^{n-x+w+\nu_2+\nu_3-1} \right]$$

Finally, following Wilcox (1979),

$$(4.6) \qquad E(\zeta|x) = \frac{1}{f(x)} \binom{n}{x} \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^{j}}{j!\, B(\nu_1+j,\nu_2,\nu_3)} \cdot$$

$$\left[\sum_{w=0}^{\infty} \binom{x}{w} B(w+\nu_2,\ n-x+\nu_3)\, B(x-w+\nu_1+j+1,\ n-x+w+\nu_2+\nu_3) \right]$$

The appealing feature of (4.1) is that unless it reduces to a Dirchlet distribution, ζ and β are correlated if the distributions of ζ and γ are assumed to be continuous. The proof of this statement follows from a result given by Darroch and Ratcliff (1971). In particular, as a special case of their theorem 2, if the probability density function of ζ and γ is continuous, the independence of ζ and β implies that ζ and γ have a Dirchlet distribution.

Numerical illustration. Data collected by the Maryland State Department of Education is used to illustrate the modified Dirichlet-multinomial model. In particular, we use the test results on students taking a preliminary form of a proficiency test in mathematics. The test consisted of thirty skills with three items per skill for a total of 90 items on the test.

We could use the information on all three items associated with each skill to obtain an averaged estimate of ζ and β for each examinee, the average being defined in the sense described by Harris and Pearlman (1978). However, since we merely want to illustrate the calculations involved in applying the model, we simply ignore the information on the third item.

For a specific examinee, we summarize the observed responses as shown in Table 2 where a 1 designates a correct and a 0 an incorrect response.

Table 2

Observed Frequencies for an Examinee

		Item 2	
		1	0
Item 1	1	x_{11}	x_{10}
	0	x_{01}	x_{00}

For example, x_{10} is the number of items the examinee is correct on the first item of an item pair and incorrect on the second.

Following Wilcox (1977) we estimate ζ with

$$\hat{\zeta} = 1 - \left(\frac{x_{01} + x_{00}}{x_{00}}\right)\left(\frac{x_{10} + x_{00}}{n}\right)$$

If $x_{00}=0$ we set $\hat{\zeta}$ equal to x_{11}/n and if $\hat{\zeta}<0$ we estimate ζ to be zero. As for β, we use $\hat{\beta} = \dfrac{x_{10}}{x_{10} + x_{00}}$

If $x_{10} + x_{00} = 0$ we set $\hat{\beta} = .25$. If $\hat{\beta} \geq .5$ we estimate β to be .5. We note that here, β represents the probability of guessing the first item in the item pair; the probability of guessing for the second item does not enter into the calculations.

The values of ζ and β were estimated using the test results on 2,000 examinees randomly sampled from the total number of examinees available. The 2,000 estimates were then used to compute the first three sample moments of ζ which were found to be .652, .496 and .405, respectively.

Since the marginal distribution of ζ is non-central beta, we can use the methods previously described to estimate ν_1, $\nu_2+\nu_3$ and λ where $\nu_2+\nu_3$ corresponds to the parameter s in section 2. The estimates are 1.2231, .83942 and .5, respectively. Next we computed the first sample moment of γ which was .1287. Since γ is assumed to have a non-central beta distribution, it follows that the mean of γ is

$$\mu_\gamma = 1-(\nu_1+\nu_3)e^{-\lambda} \int_0^1 t^{\nu_1+\nu_2+\nu_3-1} e^{\lambda} \, dt$$

Substituting .1287 for μ_γ, 1.223 for ν_1, 2.062 for $\nu_1+\nu_2+\nu_3$ and .5 for λ and solving for ν_3 yields $\hat{\nu}_3=1.279$. Thus, estimates of ν_1, ν_2, ν_3 and λ are $\hat{\nu}_1=1.2231$, $\hat{\nu}_2=.83942 - .1279=.7115$, $\hat{\nu}_3=.1279$ and $\hat{\lambda}=.5$.

An alternative extension. For a specific examinee and a randomly chosen item, let $\alpha=\Pr$ (incorrect response | examinee knows). We conclude this section by indicating that, to a certain extent, the Dirichlet-multinomial model can be extended to include the possibility of $\alpha>0$. If we allow $\alpha>0$, an examinee's percent correct true score is $\theta = (1-\alpha)\zeta+\beta(1-\zeta)$. Let $\gamma_1=\beta(1-\alpha)-\alpha\zeta$ in which case $\theta=\zeta+\gamma_1$. As long as $\beta>\alpha$, we have that $0 \le \zeta \le 1$, $0 \le \gamma_1 < 1$ and $0 \le \zeta + \gamma_1 \le 1$. Thus, it is theoretically permissible to assume ζ and γ_1 have a bivariate Dirichlet distribution, or more generally, their joint distribution is given by (4.1). Moreover, the parameters of the model can be estimated in essentially the same manner as outlined above.

References

Abramowitz, M., & Stegun, I. A. (Eds.) Handbook of mathematical functions. National Bureau of Standards, Applied Mathematics Series, Washington, D.C.: U.S. Government Printing Office, 1972, 55.

Antelman, G. R. Interrelated Bernoulli processes. Journal of the American Statistical Association, 1972, 67, 831-841.

Bardwell, G. E., & Crow, E. L. A two-parameter family of hyper-Poisson distributions. Journal of the American Statistical Association, 1964, 59, 133-141.

Chao, M. T., & Strawderman, W. E. Negative moments of positive random variables. Journal of the American Statistical Association, 1972, 67, 429-431.

Chernoff, H. The scoring of multiple choice questionnaires. Annals of Mathematical Statistics, 1962, 33, 375-393.

Connor, R. J., & Mosimann, J. E. Concepts of independence for proportions with a generalization of the Dirichlet distribution. Journal of the American Statistical Association, 1969, 64, 194-206.

Darroch, J. N., & Ratcliff, D. A characterization of the Dirichlet distribution. Journal of the American Statistical Association, 1971, 66, 641-643.

Duncan, George T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

Frary, R. B. Elimination of the guessing component of multiple-choice test scores: Effect on reliability and validity. Educational and Psychological Measurement, 1969, 29, 665-680.

Gupta, R. C. Modified power series distribution and some of its applications. Sankhya, 1974, Series B, 36, 288-298.

Hamilton, C. H. Bias and error in multiple-choice tests. Psychometrika, 1950, 15, 151-168.

Harris, C. W., & Pearlman, A. P. An index for a domain of completion or short answer items. Journal of Educational Statistics, 1978, 3, 285-304.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

References Cont.

IMSL Library 1, Volume II. Houston: International Mathematical and Statistical Libraries, 1975.

Ishii, G., & Hayakawa, R. On the compound binomial distribution. *Annals of the Institute of Statistical Mathematics*, 1960, *12*, 69-80.

James, I. R. Multivariate distributions which have beta conditional distributions. *Journal of the American Statistical Association*, 1975, *70*, 681-684.

Lord, F. M. A strong true-score theory, with applications. *Psychometrika*, 1965, *30*, 239-270.

Morrison, D. G., & Brockway, G. A modified beta-binomial model with applications to multiple choice and taste tests. *Psychometrika*, 1979, *44*, 427-442.

Noack, A. A class of random variables with discrete distribution. *Annals of Mathematical Statistics*, 1950, *21*, 127-132.

Patil, G. P. Certain properties of the generalized power series distribution. *Annals of the Institute of Statistical Mathematics*, 1962, *14*, 179-182.

Seber, G. A. The non-central chi-squared and beta distributions. *Biometrika*, 1963, *50*, 542-544.

Wilcox, R. R. Estimating the likelihood of a false-positive and false-negative decision with a mastery test: An empirical Bayes approach. *Journal of Educational Statistics*, 1977, *2*, 289-307.

Wilcox, R. Achievement tests and latent structure models. *British Journal of Mathematical and Statistical Psychology*, 1979, *32*, 61-71.

Wilcox, R. An approach to measuring the achievement or proficiency of an examinee. *Applied Psychological Measurement*, 1980, In Press (a).

Wilcox, R. Determining the length of a criterion-referenced test. *Applied Psychological Measurement*, 1980, to appear (b).

Wilcox, R. R. *Toward better approximations of the true score distribution*. Center for the Study of Evaluation, University of California, Los Angeles, 1980 (c).

Wishart, J. A note on the distribution of the correlation ratio. *Biometrika*, 1932, *24*, 441-456.

UNIVERSITY OF CALIFORNIA, LOS ANGELES

BERKELEY • DAVIS • IRVINE • LOS ANGELES • RIVERSIDE • SAN DIEGO • SAN FRANCISCO • SANTA BARBARA • SANTA CRUZ

CENTER FOR THE STUDY OF EVALUATION
UCLA GRADUATE SCHOOL OF EDUCATION
LOS ANGELES, CALIFORNIA 90024

March 28, 1980

W. Scott Gehman
Editor
Educational and Psychological
Measurement
Box 6907, College Station
Durham, NC 27708

Dear Dr. Gehman:

Please consider the enclosed manuscript "An extension of the Dirichlet-multinomial model that allows true score and guessing to be correlated" for publication in EPM.

Thank you very much.

Sincerely,

Rand R. Wilcox
Senior Research Associate

RRW/kr

Enclosure

SOME EMPIRICAL AND THEORETICAL RESULTS ON AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE

Rand R. Wilcox

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California . Los Angeles

and the

DEPARTMENT OF PSYCHOLOGY
University of Southern California

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

ABSTRACT

Wilcox (1980a) proposed a model for an answer-until-correct scoring procedure that solves various measurement problems. The purpose of this paper is to empirically check an implication of the model, and to propose and investigate some strong true-score models. One of the strong true-score models assumes the probability of guessing the correct response to an item is a strictly increasing function of an examinee's ability level, and the model gives a reasonable fit to the data. The paper illustrates that this new model is easily applied to situations where the beta-binomial model is typically used. The other models, including the Dirichlet-multinomial model, proved to be unsatisfactory. Finally, potential difficulties with the new model are discussed, and possible directions for future research are described.

1. INTRODUCTION

Wilcox (1980a) proposed a model for an answer-until-correct scoring procedure that solves various measurement problems. In particular, it can be used to test whether guessing is at random, to measure how "far away" guessing is from being random, and to correct for guessing without assuming guessing is at random. More recently, Wilcox (1980b) described six other measurement problems that the model can solve. One problem was to empirically determine the minimum number of distractors needed on a multiple-choice test item. Another can be described as follows: for a randomly selected examinee, let ε be the expected number of items on an n-item test for which we correctly determine whether the examinee knows the correct response. How many examinees do we need to sample so that there is a reasonably high probability of correctly determining whether ε has a value above or below some known constant.

Two types of guessing were considered in Wilcox (1980a). The first, or Type I guessing, refers to situations where we have a population of examinees and a single item. For a randomly sampled examinee, the probability of guessing is defined to be the Pr(correct response | examinee does not know). Type II guessing is defined in terms of a single examinee and a domain of items. In particular, it is the Pr(correct response | examinee does not know) for a randomly selected item.

In Wilcox (1980a), it was assumed that an examinee either knows the correct response and answers the item correctly, or the examinee can eliminate at most t-2 distractors where t is the number of distractors on the item. According to an answer until-correct scoring procedure, examinees choose distractors until the correct response is identified.

Assume Type I guessing, let ζ be the proportion of examinees who know the item, and let $\zeta_i (i=0,\ldots,t-2)$ be the proportion of examinees who can eliminate i distractors. Following Horst (1933), we also assume that examinees who do not know guess at random from among the distractors they cannot eliminate. Thus, the probability that a randomly chosen examinee chooses the correct alternative on the first attempt is

$$p_1 = \zeta + \sum_{i=0}^{t-2} \zeta_i/(t-i) \qquad (1.1)$$

The probability of giving the correct response on the ith attempt is

$$p_i = \sum_{j=0}^{t-i} \zeta_j/(t-j) \qquad (i=2,\ldots,t) \qquad (1.2)$$

In order for the model to hold we must have

$$p_1 \geq p_2 \geq \cdots \geq p_t \qquad (1.3)$$

Equation (1.3) can be empirically checked (Robertson, 1978). Moreover, maximum likelihood estimates of the p_i's are easily obtained when (1.3) is assumed by applying the "pool-adjacent-violators" algorithm (e.g., Barlow, et al., 1972, pp. 13-18) which in turn yields maximum likelihood estimates of the ζ's. In particular, if in a random sample of N examinees, x_1 examinees choose the correct alternative on the first try, and x_2 examinees choose the correct alternative on the second try, then

$$\begin{aligned} \hat{\zeta} &= (x_1 - x_2)/N, & x_1 \geq x_2 \\ &= 0, & x_1 < x_2 \end{aligned} \qquad (1.4)$$

is a maximum likelihood estimate of ζ. (For alternative methods of scoring and analyzing answer-until-correct tests, see Dalrymple-Alford, 1970; Brown, 1965.)

There are two main goals to this paper. The first is to empirically check the assumption in equation (1.3) for a reasonably large number of items, and the second is to propose and to empirically investigate some strong true-score models based on an answer-until-correct scoring procedure. We note that the importance of strong true-score model has long been established (e.g., Keats and Lord, 1962; Lord, 1965, 1969; Lord and Novick, 1968), and more recently they have played an important role in the realm of criterion-referenced testing (e.g., Huynh, 1976, 1980; Wilcox, 1977).

2. EMPIRICAL TESTS OF EQUATION (1.3)

As noted above, our first goal is to empirically determine whether equation (1.3) is reasonable when an answer-until-correct scoring procedure is used. To do this, we used test results on 620 students enrolled in an undergraduate psychology course. Each student took three tests during the semester. The first two tests had 37 and 40 items respectively, and the final examination had 40 items. All three tests had four forms and all items had $t=5$ distractors. Each form consisted of the same items, but they were presented in a different order. Using results in Robertson (1978), a test of equation (1.3) was made for all 117 items. Each item was tested four times according to which test form it was on. Thus, a total of 468 tests were made.

At the .01 level of significance, the null hypothesis was rejected about 5.8 percent of the time. For a little over half of the tests it was unnecessary to apply Robertson's procedure because the sample estimates of the p_i's already satisfied the inequality. When Robertson's test was applied, the results were usually highly nonsignificant. The observed test scores indicate that there were two items during each testing period (six items in all) which did not satisfy equation (1.3).

Table 1 shows the observed scores on one of the items on the final examination that appears not to satisfy (1.3). On Form 2, for example, 35 examinees chose the correct response on their third attempt of the item. For all four forms, Robertson's test was highly significant. The striking freature of this item is the large number of examinees who chose the correct response on their last attempt. One possible explanation is that examinees had misinformation relevant to the question being asked, and so they eliminated the correct response from consideration. Unfortunately, there was no way to verify this.

Several of the items for which the null hypothesis was rejected had a response pattern similar to the one shown in Table 1. That is, the correct response was usually chosen last. In another instance where the null hypothesis was rejected, the observed frequencies corresponding to the number of attempts were 20, 49, 33, 30, and 18, respectively.

3. STRONG TRUE-SCORE MODELS

Next we consider the problem of finding a strong true score model that can be used in conjunction with an answer-until-correct scoring procedure. In contrast to the previous section, only Type II guessing is considered. We begin by considering a single examinee responding to items that represent a particular item pool. Let τ be the proportion of items the examinee knows, and let $\tau_i (i=0,\ldots,t-2)$ be the proportion of items for which the examinee can eliminate i distractors when the correct response is not known. Finally let $\theta_j (j=1,\ldots,5)$ be the probability of choosing the correct response on the jth attempt of a randomly selected item. The situation is essentially the same as in the previous section, but the roles of items and examinees are interchanged.

For future reference, we note that

$$\theta_1 = \tau + \sum_{i=0}^{t-2} \tau_i/(t-i) \tag{3.1}$$

and

$$\theta_i = \sum_{j=0}^{t-i} \tau_j/(t-j) \qquad (i=\zeta,\ldots,t) \tag{3.2}$$

Let y_1 and y_2 be the number of items on an n-item test for which the examinee chooses the correct response on the first and second attempt, respectively. In Wilcox (1980a) it was assumed that the joint conditional probability function of y_1 and y_2 is given by

$$f(y_1, y_2 \mid \theta_1, \theta_2) = \frac{n!\theta_1^{y_1}\ \theta_2^{y_2}\ (1-\theta_1-\theta_2)^{n-y_1-y_2}}{y_1!\ y_2!\ (n-y_1-y_2)!} \tag{3.3}$$

This implies that $f(y_1|\theta_1)$ is a binomial probability function which, in mental test theory, has certain theoretical disadvantages. In practice, however, this assumption frequently gives good results. A recent discussion of the issues can be found in Wilcox (1981).

Note that for the model to hold, we must have

$$\theta_1 \geq \theta_2 \geq \cdots \geq \theta_t$$

and so a maximum likelihood estimate of τ is

$$\begin{aligned} \hat{\tau} &= (y_1-y_2)/n, & y_1 &\geq y_2 \\ &= 0, & y_1 &< y_2 \end{aligned}$$

The goal in this section is to consider how we might extend (3.3) to a population of examinees. Wilcox (1980a) suggests that for a population of examinees, we assume the joint density of θ_1 and θ_2, or the joint density of τ and θ_2 belongs to the Dirichlet family. For the former case, the joint density is given by

$$g(\theta_1, \theta_2) = \frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1)\Gamma(\nu_2)\Gamma(\nu_3)} \theta_1^{\nu_1-1} \theta_2^{\nu_2-1} (1-\theta_1-\theta_2)^{\nu_3-1} \qquad (3.4)$$

where ν_1, ν_2, $\nu_3 > 0$ are unknown parameters. In the latter case we simply replace θ_1 with τ in equation (3.4). The motivation for (3.4) is that it is the bivariate analog of the beta density which has proven to be useful in many situations in mental test theory (Wilcox, 1981).

Empirical Results on the Dirichlet-multinomial Model

For the reasons given above, we began by assuming (3.4), and we then tried to fit the Dirichlet-multinomial model to the final examination test scores previously described. From results in section 2, two of the forty items appear not to satisfy the assumptions made under our answer-until-correct scoring procedure and Type I guessing, and so they were eliminated. The observed marginal distribution of y_1 and y_2 for the remaining 38 items is shown in Table 2.

It is known that when (3.4) is assumed, the marginal density of θ_1 is beta with parameters ν_1 and $\nu_2+\nu_3$. We tried fitting the observed y_1's to a beta-binomial probability function (e.g., Keats and Lord, 1962). The estimates of ν_1 and $\nu_2+\nu_3$ were 8.645 and 8.2, respectively. The expected frequencies under the model are shown in Table 2. A visual inspection of Table 2 suggests that the beta-binomial model gives a reasonable fit to the data, and a chi-square goodness-of-fit test (Cochran, 1954) confirms this.

Next we consider the observed frequencies corresponding to y_2. If (3.4) is assumed, then the marginal probability function of y_2 is

$$f(y_2) = \binom{n}{y_2} \frac{\Gamma(\nu_1+\nu_2+\nu_3)\,\Gamma(y_2+\nu_2)\Gamma(n+\nu_1+\nu_3-y_2)}{\Gamma(\nu_2)\Gamma(\nu_1+\nu_3)\,\Gamma(n+\nu_1+\nu_2+\nu_3)} \tag{3.6}$$

i.e., a beta-binomial density with parameters ν_2 and $\nu_1+\nu_3$. The estimate of ν_2 was 25.6, and the estimate of $\nu_1+\nu_3$ was 101.6. Again a good fit to the data was obtained. However, the estimates of ν_1, ν_2 and $\nu_2+\nu_3$ imply that ν_3 must be negative. But the Dirichlet-multinomial model assumes $\nu_i>0$ $(i=1,2,3)$. We tried instead to estimate the ν_i's as described in Mosimann (1962). This yielded $\hat{\nu}_1=6.08$, $\hat{\nu}_2=2.37$ and $\hat{\nu}_3=3.39$. We now have admissible estimates of the ν_i's, but the fit to data is no longer satisfactory. Evidently, some other model must be used to explain the observed scores.

Before describing a model that gives a reasonably good fit to the data, we might mention two other models that were considered but which gave unsatisfactory results. The first was a negative-multinomial model (e.g., Sibuya et al., 1964), and the second was a compound negative multinomial model also known as the multivariate inverse Polya-Eggenberger distribution (Mosimann, 1963; Sibuya, 1980; Sibuya and Shimizu, 1980; Janardan and Patil, 1971).

A New Strong True-Score Model

Since a beta-binomial model gives a good fit to the observed marginal distribution of y_1, we decided to assume (3.3) holds and that θ_1 has a beta density with parameters 8.645 and 8.2. The problem is to find a reasonable relationship between θ_1 and θ_2 that accounts for the observed marginal density of y_2. As noted above, the Dirichlet-multinomial model, as well as two other models, is unsatisfactory for accomplishing this goal.

Our common sense notion is that as τ increases, the probability of guessing the correct response will also increase. For instance, an examinee with a value for τ close to one might have more partial information than an examinee for whom τ is small. That is, examinees with τ close to one might be able to eliminate more distractors when they do not know as opposed to examinees for whom τ is close to zero. We note that Molenaar (in press) has also argued for this point of view. Let's assume for the moment that this is true, and consider how we might express this relationship.

After looking at the data, we decided to express the assumed relationship between τ and guessing in terms of the conditional distribution of y_2 given y_1. First note that for a specific examinee

$$f(y_2|y_1, \theta_1, \theta_2) = \binom{n-y_1}{y_2}\left(\frac{\theta_2}{1-\theta_1}\right)^{y_2}\left(1-\frac{\theta_2}{1-\theta_1}\right)^{n-y_1-y_2} \tag{3.7}$$

For notational convenience, let $\xi=\theta_2/(1-\theta_1)$. Our assumption about τ and guessing indicates that for the population of examinees, ξ is an increasing function of θ_1. Since $0\leq\theta_1\leq1$, what we need is an increasing function that maps the closed unit interval into a subset of itself. One way to do this is to use a linear function of a cumulative distribution defined on $[0,1]$. The beta distribution is the best known distribution with this property, and so we decided to consider it for the problem at hand. Accordingly, we assume that for the population of examinees, $E(\xi|\theta_1)$ is given by

$$\xi(\theta_1) = c\int_0^{\theta_1} \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} u^{r-1}(1-u)^{s-1}\,du + e \tag{3.8}$$

where c, e, r, and s are unknown positive constants to be determined, and where $0<c+e<1$.

Henceforth, we assume θ_2 is completely determined by θ_1 according to equation (3.8). That is, for a specific examinee, $\theta_2=(1-\theta_1)\xi(\theta_1)$. This is, no doubt, an over simplification of reality, but we want to avoid deriving a model so mathematically complex that it cannot be applied. As it turns out, equation (3.8) gives a reasonably good fit to the data.

Next we determined c, e, r and s in the manner described in the appendix. The results were $\hat{c}=.25$, $\hat{e}=.25$, $\hat{r}=1.776$ and $\hat{s}=2.279$.

As a partial check on the model, we decided to compare the expected observed scores of y_2 to the values actually observed. To do this, we need an expression for the marginal distribution of y_2 assuming equations (3.8) and (3.3) hold, and that θ_1 has a beta density with parameters 8.645 and 8.2. Writing $\xi(\theta_1)$ simply as ξ and since $\theta_2=\xi(1-\theta_1)$ equation (3.3) can be written as

$$f(y_1,y_2|\theta_1) = n!\theta_1^{y_1}[\xi(1-\theta_1)]^{y_2}[1-\theta_1-\xi(1-\theta_2)]^{n-y_1-y_2} \tag{3.9}$$

and

$$f(y_2|\theta_1) = \frac{n!\xi^{y_2}(1-\theta_1)^{y_2}[1-\xi(1-\theta_1)]^{n-y_2}}{y_2!(n-y_2)!} \tag{3.10}$$

Substituting (3.8) into (3.10), multiplying by $g(\theta_1)$, and integrating out $g(\theta_1)$ yields the maginal density of y_2. Symbolically,

$$f(y_2) = \int_0^1 f(y_2|\theta_1)\, g(\theta_1)d\theta_1 \tag{3.11}$$

where, from previous results, we assume

$$g(\theta_1) = \frac{\Gamma(16.845)}{\Gamma(8.645)\Gamma(8.2)} \theta_1^{7.645} (1-\theta_1)^{7.2} \qquad (3.12)$$

and $f(y_2|\theta_1)$ is given by (3.10). Since ξ is a function of θ_1, it is difficult to find a closed form expression for (3.12). However, for practical purposes, this is not a serious problem since the integration is easily accomplished using numerical quadrature techniques. We used the IBM (1971) subroutine DQG32. For those who do not have access to this subroutine, the necessary formulas can be found in Stroud and Secrest (1966). The expected scores of y_2 based on (3.12) are shown in the last column of Table 2. The usual chi-square statistic was found to be 22.9. With 12 degrees of freedom the level of significance is between .025 and .05. (Note that e is assumed known, as is explained in the appendix, and so (3.11) has three unknown paramenters since (3.12) is assumed.)

We observe that the estimates of ξ corresponding to y_1=4, 7 and 33 are based on a relatively small number of examinees. In fact, for y_1=4 there is only one examinee and the same is true for y_1=33. Thus, it might be that $\hat{\xi}$ is unusually spurious at these points, and this would explain why we get estimates of ξ that seem to be relatively inconsistent with the notion that ξ is a strictly increasing function of θ_1. (See Table A2 in the appendix.)

It is interesting that if we ignore the estimates of ξ at these points, we get $\hat{c}$=.33, $\hat{r}$ .88 and $\hat{s}$=.909 with e still equal to .25. In this case the value of the chi-square statistic is 15.63 and the level of significance is between .05 and .1. In either case, we get a reasonable approximation to the data. Note, however, that if we assume random guessing, i.e., $\xi = (t-1)^{-1} = .25$, as is frequently done, we get a very poor fit to the data.

Next we applied the model to the observed scores on the second test taken during the semester. We used the same 620 examinees. Again, two of the forty items did not satisfy (1.3), and so they were eliminated. The parameters of our strong true-score model were estimated and found to be very similar to the estimated values based on the final examination. Also, we again got a reasonable fit to the data.

Before concluding this section we note that the above results suggest we estimate τ with $\hat{\tau}=\hat{\theta}_1-\hat{\theta}_2=\hat{\theta}_1-(1-\hat{\theta}_1)\xi$. If we arbitrarily set $\xi=(t-1)^{-1}$, we get the usual correction for guessing formula score.

4. SOME APPLICATIONS TO MASTERY TESTS

In many instances it is a simple matter to extend existing applications of the beta-binomial model to the model described in section 3. By way of illustration, we consider two problems that occur with mastery tests.

A frequent goal of a mastery or criterion-referenced test is to sort examinees into one of two mutually exclusive groups. In many instances these groups are defined according to whether an examinee's true score τ is above or below some known constant, say τ_0. In the context of an answer-until-correct scoring procedure, we decide that $\tau \geq \tau_0$ if for the examinee being tested, $(y_1-y_2)/n \geq \tau_0$; otherwise the decision $\tau < \tau_0$ is made.

For a randomly selected examinee, the probability of making a correct decision about whether τ is above or below τ_0 is given by

$$Pr(y_1-y_2 \geq [n\tau_0],\ \tau \geq \tau_0)+Pr(y_1-y_2<[n\tau_0],\ \tau<\tau_0) \qquad (4.1)$$

where $[n\tau_0]$ is the smallest integer greater than or equal to $n\tau_0$. But (4.1) is equal to

$$\Sigma\int_{\theta_0}^{1} f(y_1,y_2|\theta_1)g(\theta_1)d\theta_1 + \Sigma\int_{0}^{\theta_0} f(y_1,y_2|\theta_1)g(\theta_1)d\theta_1$$

where the first summation is over all (y_1, y_2) such that $y_1-y_2 \geq [n\tau_0]$, the second summation is over all (y_1,y_2) such that $y_1-y_2<[n\tau_0]$ and θ_0 is the value of θ_1 such that $\theta_1-(1-\theta_1)\xi=\tau_0$. Thus, the probability of a correct decision can be determined once (3.8) is estimated.

Another approach to characterizing mastery tests is the single administration estimate of the proportion of agreement. We are given the observed scores of N examinees, and we want to estimate the probability that a randomly selected examinee would be classified in the same manner if he/she took two randomly parallel tests.

Let z_1 and z_2 be the observed scores corresponding to y_1 and y_2 for an examinee who takes a randomly parallel test. Proceeding in a manner similar to Huynh (1976), assume the density $f(z_1, z_2|\theta_1)$ has the same form as $f(y_1,y_2|\theta_1)$ which is given by (3.9). Thus, after making the appropriate independence assumption, the joint density of y_1, y_2, z_1, and z_2 is

$$f(y_1,y_2,z_1,z_2) = \int_0^1 f(y_1,y_2|\theta_1)\ f(z_1,z_2|\theta_1)\ g(\theta_1)d\theta_1,$$

which can be evaluated with IBM subroutine DQG32. The proportion of agreement is

$$\Sigma f(y_1,y_2,z_1,z_2)$$

where the summation is over all points where both y_1-y_2 and z_1-z_2 are greater than or equal to $n\tau_0$, or when both are less than or equal to $n\tau_0$.

5. DIRECTIONS FOR FUTURE RESEARCH

We briefly describe some of the problems that might occur when using the strong true-score model proposed in section 3.

First, the assumption that the marginal probability function of y_1 belongs to the beta-binomial family has yielded good results to various measurement problems when applied to real data (e.g., Gross and Shulman, 1980; Subkoviak, 1978; Keats and Lord, 1962; Lord, 1965). However, as might be expected, this is not always the case. Keats (1964a) reports a data set for which the beta-binomial model gives a poor fit, and Keats (1964b) reports several other data sets for which the model gives unsatisfactory results. Accordingly, we briefly outline solutions that might be considered when the beta-binomial model is unsatisfactory. The details are left for future investigations.

First we note that when trying to find a probability function that gives a good fit to data, three of the best known and most frequently employed distributions are the binomial, Poisson and negative binomial. Of course, the Poisson distribution usually gives good results when applied to situations where a particular event occurs infrequently. Also, the negative binomial distribution is often the first choice when it is believed that the Poisson distribution might be inadequate (Johnson and Kotz, 1969, p. 125).

Suppose we replace (3.3) with the assumption that for a particular examinee, the probability of $z=n-y_1$ is

$$f(z|\gamma) = e^{-\gamma}\,\gamma/z! \qquad (z=0,1,\ldots) \qquad (5.1)$$

i.e., a Poisson density with parameter γ. If we also assume γ has a gamma distribution for the population of examinees, the marginal distribution of z is negative binomial given by

$$f(z) = \binom{\alpha+z-1}{\alpha-1} \left(\frac{\beta}{\beta+1}\right) \left(\frac{1}{\beta+1}\right)^{\alpha} \qquad (5.2)$$

where α and β are unknown parameters. As noted in Wilcox (1981), this distribution gives a reasonable fit to the data reported by Keats (1964a) while the beta-binomial model does not. We also note that Johnson and Kotz (1969) list several techniques for estimating the parameters in (5.2).

One problem is how to represent the joint distribution of y_2 and z. A mathematically convenient approach is to assume y_2 is also Poisson and that z and y_2 are (conditionally) independent. To allow z and y_2 to be correlated, we might use the bivariate Poisson distribution derived by Holgate (1964).

In principle, at least, a gamma-Poisson model could be applied, and an estimate of ξ could be derived. However, if test scores are highly skewed, as they are for the data in Keats (1964a), we might get poor estimates of ξ for examinees with low ability because there are so few examinees with low ability. Hopefully the seriousness of this problem will be investigated sometime in the future.

Rather than assume γ has a gamma distribution, we might assume it has a gamma product ratio distribution in which case the marginal probability function of z is

$$f(z) = \frac{\Gamma(\alpha+\omega)\Gamma(\beta+\omega)}{\Gamma(\alpha+\beta+\omega)\Gamma(\omega)} \cdot \frac{(\alpha)_z\,(\beta)_z}{z!\,(\alpha+\beta+\omega)_z} \qquad (5.3)$$

where α, β, $\omega>0$ and

$$(a)_z=1, \quad z=0$$

$$= a(a+1)\ldots(a+z-1), \quad z=1,2,\ldots$$

(Sibuya, 1979). The distribution (5.3) is known as the inverse Polya-Eggenberger, the generalized Waring, and the negative binomial beta. The last term is sometimes used because if $f(z|\gamma)$ is negative binomial with parameters α and p, and if p is beta with parameters ω and β, the marginal distribution of z is (5.3).

The rth factorial moment of (5.3) is

$$\mu_r = E(z^{(r)})$$

$$= (\alpha)_r(\beta)_r/(\omega-1)^{(r)}$$

where $a^{(r)} = a(a-1)\ldots(a-r+1)$.

We can estimate ω by the method of moments by noting that

$$\left(\frac{\mu_2}{\mu_1} - \mu_1\right)\omega - \frac{2\mu_2}{\mu_1} = -\mu_1 + \alpha + \beta + 1$$

and

$$\left(\frac{\mu_3}{\mu_1} - \mu_1\right)\omega - \frac{3\mu_3}{\mu_2} = -\mu_1 + 2\alpha + 2\beta + 4$$

The values of α and β can then be determined via the estimate of μ_1 and μ_2.

We note that Irwin (1968) reports some real data for which (5.3) improves upon the fit obtained with the negative binomial, but the improvement is not overly striking.

As for the joint distribution of z and y_2, we might use the multivariate analog of (5.3). (See, for example, Sibuya, 1980.) Once α is estimated, results in Mosimann (1963) can be used to estimate the remaining parameters.

6. CONCLUDING REMARKS

There are two main points to this paper. First, Wilcox (1980a) made certain assumptions about how examinees behave when responding to test items according to an answer-until-correct scoring procedure. These assumptions imply that the cell probabilities in a multinomial distribution must satisfy a particular set of inequalities. The data used in this study suggests that these inequalities will frequently hold.

The second point is that a strong true-score model was proposed that allows the probability of guessing the correct response to vary over a population of examinees. In particular, it was assumed that the probability of guessing correctly is a strictly increasing function of an examinees ability level. Furthermore, the model gives a reasonably good fit to our data, and it allows us to correct for guessing without assuming guessing is at random.

Finally, we have outlined some of the potential difficulties with our proposed model. Hopefully these issues will be resolved sometime in the future.

APPENDIX

The problem is to derive an estimate of the parameters in equation (3.8). To motivate our solution, we first rederive the estimate of the true score distribution used by Lord and Novick (1968, chapter 23). The point is that the derivation is done in slightly different fashion than is customary. We then apply this same technique to obtain an estimate of ξ as a function of θ_1.

Suppose that on an n-item test, observed scores for a specific examinee have a probability function given by

$$f(x|\pi) = \binom{n}{x} \pi^x(1-\pi)^{n-x}$$

For a population of examinees, let $g(\pi)$ be the density of π, and suppose we want to estimate the first two moments of π. One approach to this problem is as follows: Let x_i $(i=1,\ldots,N)$ be the observed scores of N randomly sampled examinees, and let f_x be the number of examinees with an observed x where, of course, $\Sigma f_x=N$. Temporarily assume that every examinee's true score π has one of n+1 possible values, namely, $\pi_i=i/n$ $(i=0,\ldots,n)$. The observed values of x suggest that we have sampled f_x examinees with true score x/n. Thus, an estimate of the probability of choosing an examinee having true score π_i is $h(\pi_i)=f_x/N$. Since $\hat{\pi}_i=x_i/n$ is an unbiased estimate of π for the ith examinee, an estimate of $E(\pi)$ is

$$\sum_{i=0}^{n} \hat{\pi}_i \, h(\pi_i) = \sum_{x=0}^{n} \frac{x}{n} \frac{f_x}{N} = \sum_{i=1}^{N} \frac{x_i}{nN} .$$

Since $n^{-1}(n-1)^{-1}x(x-1)$ is an unbiased estimate of π^2, this suggests, for similar reasons, that we estimate $E(\pi^2)$ with

$$N^{-1} \sum_{i=1}^{N} \frac{x_i^2-x_i}{n(n-1)} .$$

These estimates of $E(\pi)$ and $E(\pi^2)$ are the same as the ones derived in Lord and Novick (1968). If we now assume $g(\pi)$ belongs to the beta family, we have their estimate of the true score distribution.

In this paper we assume that θ_1 completely determines ξ, and that ξ is given by (3.8). Temporarily assume that θ_1 is discrete, and that its possible values are i/n $(i=0,\ldots,n)$. Suppose we want to estimate the value of ξ for the possible values of θ_1. We do this as follows.

For notational convenience let $y=y_1$, and suppose f_y examinees get y items correct on their first attempt of an item. Thus, we would estimate that f_y examinees have $\theta_1=y/n$. Let $h(y_2|y_1)$ be the number of examinees who get y_2 corrects on the second try of an item given that there were y items for which the examinee chose the correct response on the first try. Finally, let

$$\hat{\xi} = \sum_{y_2} \frac{y_2 h(y_2|y)}{(n-y)h} .$$

where $h=\sum_{y_2} h(y_2|y)$. Then $\hat{\xi}$ is an estimate of ξ when $\theta_1=y/n$.

We illustrate the calculations using a specific case from the data reported in the paper. Consider $y=11$. The corresponding y_2 values for which $h(y_2|y)$ is positive are $y_2=8,9,10,11,13$ and 14. The frequencies (the values of $h(y_2|y)$) were 4,5,1,1,1,2, respectively. Thus, $h=14$. Since there are $n=38$ items, we would estimate $\xi(11/38)$ to be $139/((38-11)(14))=.36$.

Table A1 shows the estimates of $\xi(\theta_1)$ for the final examination test scores used in the paper. The values of $\hat{\xi}$ suggest that ξ is indeed an increasing function of θ_1, but occasionally ξ decreases. According, we applied the pool-adjacent-violators algorithm (Barlow, et al., 1972, pp. 13-15) to estimate ξ under the assumption that it is a nondecreasing function of θ_1. The results are reported in Table A1 as $\tilde{\xi}$.

Since there are t=5 distractors for every item on the test, and since, for a specific examinee, ξ is the probability of a correct on the second try when the examinee is incorrect on the first try, the values of $\tilde{\xi}$ suggest that examinees with low ability are guessing approximately at random. We decided in advance to set $e=(t-1)^{-1}=.25$, and the data suggests that this is reasonable. Based on Table A1, we also assume that the upper value of ξ is .50, and so we set $c=.50-e=.25$.

There remains the problem of estimating r and s. First, since ξ is assumed to be a strictly increasing function of θ_1, we cannot use the same estimates of ξ for two distinct values of θ_1. Suppose θ_{1i} $(i=1,\ldots,m)$ are m points where the estimate of ξ (the value of $\tilde{\xi}$) is the same. For the purpose of estimating r and s, we replace the points θ_{1i} $(i=1,\ldots,m)$ with $m^{-1}\Sigma\theta_{1i}$. For example, in Table A1, we have that $\tilde{\xi}=.305$ at $\theta_1=.24$ and .26. Thus, instead of using the two points $\theta_1=.24$ and .26, we assume $\xi=.305$ at $\theta_1=.25$, and that a value of ξ at $\theta_1=.24$ and .26 is not available. The resulting values of θ_1 and the corresponding values of $\tilde{\xi}$ are shown in Table A2.

Next set $\eta=(\xi-\overset{e}{\cancel{.25}})/\overset{c}{\cancel{.25}}$ and note that $\eta=\int_0^{\theta_1}\frac{\Gamma(r+s)}{\Gamma(r)\ (s)}\overset{\mu}{\cancel{\omega}}^{r-1}(1-\overset{\mu}{\cancel{\omega}})^{s-1}du.$ The value of η corresponding to the values θ_1 are summarized in Table A2. They give us a step function approximation to an assumed cumulative beta distribution. Thus, by calculating the mean and variance of this step function, we can estimate r and s (e.g., Lord and Novick, 1968, chapter 23; Wilcox, 1977). For the data used here, the estimates were $\hat{r}=1.776$ and $\hat{s}=2.279$, respectively.

TABLE 1

Observed Frequencies for an Item Not Satisfying (1.3)

		Number of Attempts				
		1	2	3	4	5
	1	19	21	24	36	57
Test Form	2	16	22	35	30	51
	3	13	14	33	24	67
	4	13	15	42	34	52

TABLE 2

Observed and Expected Scores on the Final Examination

Score	Observed Frequency of y_1	Observed Frequency of y_2	Expected y_1 when y_1 is bebi (8.645, 8.2)	Expected y_2 when y_2 is bebi (25.6, 101.61)	Expected y_2 When c=e=.25 r=1.2776 s=2.279
0	0	2	.00	.37	.62
1	0	5	.02	2.67	3.66
2	0	10	.07	9.55	11.72
3	0	24	.20	23.19	26.72
4	1	34	.48	42.78	47.55
5	5	51	1.00	63.98	69.63
6	6	90	1.87	80.60	86.30
7	9	85	3.20	87.85	92.26
8	4	90	5.08	84.26	86.12
9	14	75	7.58	72.30	70.93
10	19	64	10.71	56.00	51.89
11	25	46	14.44	39.43	33.91
12	26	25	18.67	25.48	19.84
13	34	7	23.22	15.13	10.48
14	26	7	27.87	8.31	4.96
15	34	3	32.37	4.22	2.17
16	43	1	36.43	1.98	.87
17	42	1	39.79	.87	.31
18	46	0	42.22	.37	.12
19	41	0	43.54	.12	0.00
20	45	0	43.63	.06	0.00
21	46	0	42.51	0.00	0.00
22	40	0	40.25	0.00	0.00
23	38	0	36.97	0.00	0.00
24	28	0	32.94	0.00	0.00
25	25	0	28.41	0.00	0.00
26	19	0	23.66	0.00	0.00
27	27	0	18.97	0.00	0.00
28	13	0	14.59	0.00	0.00
29	11	0	10.72	0.00	0.00
30	6	0	7.48	0.00	0.00
31	4	0	4.90	0.00	0.00
32	6	0	3.00	0.00	0.00
33	1	0	1.68	0.00	0.00
34	2	0	.84	0.00	0.00
35	1	0	.37	0.00	0.00
36	0	0	.13	0.00	0.00
37	0	0	.03	0.00	0.00
38	0	0	.00	0.00	0.00

Explanation of notation, y_1 is bebi (a, b) means y_1 has a beta-binomial density with parameters a and b.

TABLE A1

y_1	$\theta_1 = y_1/n$	$\hat{\xi}$	$\tilde{\xi}$
4	.11	.23	.298
7	.18	.41	.298
8	.21	.26	.298
9	.24	.32	.305
10	.26	.29	.305
11	.29	.36	.355
12	.32	.35	.355
13	.34	.37	.37
14	.37	.39	.385
15	.39	.38	.385
16	.42	.42	.41
17	.45	.40	.41
18	.47	.40	.41
18	.47	.40	.41
19	.50	.45	.43
20	.53	.42	.43
21	.55	.43	.43
22	.58	.48	.44
23	.61	.42	.44
24	.63	.43	.44
25	.66	.43	.44
26	.68	.50	.46
27	.71	.42	.46
28	.74	.55	.50
29	.76	.45	.50
30	.79	.63	.50
31	.82	.57	.50
32	.87	.20	.50

TABLE A2

θ_1:	.17	.24	.30	.34	.38	.45	.53	.62	.70	.80
$\tilde{\xi}$:	.298	.305	.355	.37	.385	.41	.43	.44	.46	.5
η :	.19	.22	.42	.48	.53	.64	.72	.76	.84	1.00

References

Barlow, R. E., Bartholomew, D. J., Bremmer, J. M., & Brunk, H. D. (1972) Statistical inference under order restrictions. New York: Wiley.

Brown, J. (1965) Multiple response evaluation of discrimination. *The British Journal of Mathematical and Statistical Psychology*, *18*, 125-137.

Cochran, W. G. (1954) Some methods for strengthening the common X^2 tests. *Biometrics*, *10*, 417-451.

Dalrymple-Alford, E. C. (1970) A model for assessing multiple-choice test performance. *British Journal of Mathematical and Statistical Psychology*, *23*, 199-203.

Gross, A. L., & Shulman, V. (1980) The applicability of the beta-binomial model for criterion-referenced testing. *Journal of Educational Measurement*, *17*, 175-202.

Holgate, P. (1964) Estimation for the bivariate Poisson distribution. *Biometrika*, *51*, 241-244.

Horst, P. (1933) The difficulty of a multiple choice test item. *Journal of Educational Psychology*, *24*, 229-232.

Huynh, H. (1976) On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, *13*, 253-264.

Huynh, H. (1980) Statistical inference for false positive and false negative error rates in mastery testing. *Psychometrika*, *45*, 107-120.

Irwin, J. O. (1968) The generalized Waring distribution applied to accident theory. *Journal of the Royal Statistical Society*, *131*, Series A, 205.

Janardan, K. G., & Patil, G. P. (1971) The multivariate inverse Polya distribution: A model of contagion for data with multiple counts in inverse sampling. Studi di Probabilita, Statistica e Ricerca Operativa in onore di Giuseppe Pompilj, Oderisi-Gubbio, 1-15.

Johnson, N., & Kotz, S. (1969) Discrete Distributions. New York: Wiley.

Keats, J. A. (1964) Some generalizations of a theoretical distribution of mental test scores. Psychometrika, 29, 215-231.

Keats, J. A. (1964) Survey of test score data with respect to curvilinear relationships. Psychological Reports, 15, 871-874.

Keats, J. A., & Lord, F. M. (1962) A theoretical distribution for mental test scores. Psychometrika, 27, 59-72.

Lord, F. M. (1965) A strong true-score theory, with applications. Psychometrika, 30, 239-270.

Lord, F. M. (1969) Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 34, 259-299.

Lord, F. M., & Novick, M. R. (1968) Statistical theories of mental test scores. Reading, Mass.: Addison - Wesley.

Molenaar, I. W. (in press) On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology.

Mosimann, J. E. (1962) On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika, 49, 65-82.

Mosimann, J. E. (1963) On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. Biometrika, 50, 47-54.

Robertson, T. (1978) Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 73, 197-202.

Sibuya, M. (1979) Generalized hypergeometric, digamma and trigamma distributions. Annals of the Institute of Statistical Mathematics, 31, 373-390.

Sibuya, M. (1980) Multivariate digamma distribution. Annals of the Institute of Statistical Mathematics, 32, Part A, 25-36.

Sibuya, M., & Shimizu, R. (1980) Classification of the generalized hypergeometric family of distributions. The Institute of Statistical Mathematics, Research memorandum No. 192.

Sibuya, M., Yoshimura, I., & Shimizu, R. (1964) Negative multinomial distribution. Annals of the Institute of Statistical Mathematics, 16, 409-426.

Stroud, A. H., & Secrest, D. (1966) Gaussian quadrature formulas. New Jersey: Prentice-Hall.

Subkoviak, M. J. (1978) Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 15, 111-116.

Wilcox, R. R. (1977) Estimating the likelihood of false-positive and false negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics, 2, 289-307.

Wilcox, R. R. (1980) Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, in press.

Wilcox, R. R. (in press) Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement.

Wilcox, R. R. (1981) A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 6, 3-32.

Wilcox, R. R. (in press) The single administration estimate of the proportion of agreement of a proficiency test scored with a latent structure model. Educational and Psychological Measurement.

SOME NEW RESULTS ON AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE

Rand R. Wilcox

DEPARTMENT OF PSYCHOLOGY
University of Southern California
Los Angeles, California 90007

and the

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles 90024

ABSTRACT

Wilcox (1981a, 1982) proposed a method of scoring and analyzing achievement tests and achievement test items that might be used to solve various measurement problems including correcting for guessing without assuming guessing is at random. The new procedure is based on certain assumptions about how examinees behave when taking an answer-until-correct test. Certain implications of these assumptions have been empirically checked and the results suggest that Wilcox's model will frequently be reasonable. The purpose of this paper is to see whether similar results will be obtained when a different type of achievement test is used with a substantially different population of examinees. Included is a simplification of Wilcox's strong true-score model that gives a good fit to one of the data sets. The paper also notes that a knowledge or random guessing model is highly unsatisfactory when trying to explain the observed test scores. Finally, a new model for measuring misinformation is proposed and found to give good results with two of the items.

Under an answer-until-correct (AUC) scoring procedure, examinees choose alternatives on a multiple-choice test item until the correct response is identified. In the past this has been accomplished by having examinees erase a shield on an answer sheet which reveals whether the correct response was chosen. If an incorrect alternative was selected, another shield is erased, and this process continues until the examinee chooses the correct alternative.

Wilcox (1981a; 1982, in press a) proposed a method of scoring and analyzing AUC tests that solves various measurement problems. These include correcting for guessing without assuming guessing is at random, testing whether guessing is at random, measuring "how far away" guessing is from being at random, estimating the accuracy of know/don't know decisions when a conventional scoring procedure is used, and empirically determining the number of distractors needed on a multiple-choice test. Wilcox also derived a strong true-score model that allows the probability of guessing the correct response to vary over the population of examinees, and the model also allows true score and the probability of guessing to be correlated. The new model contains the beta-binomial model (Lord & Novick, 1968, chapter 23; Wilcox, 1981b) and the Morrison & Brockway (1979) model as a special case. The scoring procedure has been applied to criterion-referenced tests (Wilcox, in press b, in press c) and found to substantially reduce the problems noted by van den Brink and Koele (1980) and Wilcox (1980).

The purpose of this paper is to empirically investigate certain implications of the assumptions made by Wilcox, to suggest a new model for measuring misinformation, and to indicate a modification of Wilcox's strong true-score model that might be used in certain situations.

2. METHODS AND RESULTS

Consider a randomly sampled examinee responding to a specific test item under an AUC scoring procedure. Let p_i be the probability that the correct response is chosen on the ith attempt of the item, and suppose that examinees who do not know the correct response can eliminate at most t-2 distractors from consideration via partial information. Once the examinee eliminates as many distractors as he/she can, a response is chosen at random from among those remaining. If the randomly sampled examinee knows the correct response, it is assumed that the correct alternative is chosen on the first attempt.

If ζ is the proportion of examinees who know the correct response, and if ζ_i is the proportion who can eliminate i distractors, then the p_i's can be written as linear combinations of the ζ's. For example, if there are t=4 alternatives,

$$
\begin{aligned}
p_1 &= \zeta + \zeta_0/4 + \zeta_1/3 + \zeta_2/2 \\
p_2 &= \zeta_0/4 + \zeta_1/3 + \zeta_2/2 \\
p_3 &= \zeta_0/4 + \zeta_1/3 \\
p_4 &= \zeta_0/4
\end{aligned}
$$

and so

$$\zeta = p_1 - p_2 .$$

Thus, if N examinees are tested, and if x_i examinees are correct on their ith attempt of an item, the estimate of ζ is simply $\hat{\zeta} = (x_1-x_2)/N$. Moreover, the above results easily generalize to any t (Wilcox, 1981a), and it can be seen that

$$p_1 \geq p_2 \geq \cdots \geq p_t . \tag{2.1}$$

A test of mechanical abilities was administered to examinees in Great Britain who were approximately 14 years old. Each item required the examinee to apply some physical law in order to solve a problem. For example, one of the questions was stated as follows:

"Where can a jet plane not fly?"

The alternatives were (A) over deep water, (B) over high mountains, (C) over mountains on the moon, (D) very low, (E) 8 miles above the earth.

Results in Robertson (1978) were applied to each of the 30 items to test whether equation 2.1 might holds. The first 15 items had t=5 alternatives, and the remaining 15 had t=3. The x_i values are shown in Table 1. There were 386 examinees, but some examinees omitted certain items. For 20 of the items, Robertson's test was not necessary since the estimated p_i values were already consistent with equation 2.1. Among the remaining items two were significant at the .01 level (items 7 and 30 in Table 1), one was significant at the .05 level (item 29), and the remaining items were not significant at the .25 level.

3. THE MODEL AS A DIAGNOSTIC TOOL

When measuring achievement, particularly within an instructional setting, it would be helpful to have some method of detecting misinformation, identifying the type of misinformation being used, and when it exists, measuring how pervasive this misinformation is. Of course the teacher's judgment of how the students are behaving on a test is an integral part of diagnosing misinformation. The results reported here are intended to supplement or possibly help verify the teacher's view. Included is a modification of Wilcox's model which might be helpful in this endeavor.

As noted in the previous section, item 7 proved to be inconsistent with Wilcox's model, and the natural reaction is to try to determine why we got this result. The item was worded as follows:

A block of iron weighs 40 newtons at room temperature. When it is heated until it is red hot it gets bigger. How much will it weigh when red hot?

(A) 39 newtons, (B) 40 newtons, (C) 40.5 newtons, (D) 41 newtons, and (E) 42 newtons.

It seems reasonable that some examinees might believe that because the iron is bigger when red hot, it should weigh more. Thus, examinees will eliminate A and B from consideration and choose from among the responses C, D, and E. If the proportion of examinees acting in this manner is reasonably large, we would expect a disproportionate number of examinees requiring 4 attempts to identify the correct response, and this is consistent with the frequencies in Table 1.

For the reasons just outlined, it seems that Wilcox's model is inappropriate for item 7, and that the following model be used in its place. Let ζ be the proportion of examinees who know the correct response, and suppose that examinees who know are always correct on their first attempt. Let ζ_1 be the proportion who do not know and choose alternatives at random, and let ζ_2 be the proportion of examinees who believe that the iron weighs more when heated because it is bigger. If these three categories are the only ones to which an examinee can belong, then

$$p_1 = \zeta + \zeta_1/5 \tag{3.2}$$

$$p_2 = \zeta_1/5 \tag{3.3}$$

$$p_3 = \zeta_1/5 \tag{3.4}$$

$$p_4 = \zeta_2 + \zeta_1/5 \tag{3.5}$$

and

$$p_5 = \zeta_1/5 \tag{3.6}$$

Note that this model is similar to the misinformation used by Duncan (1974).

An obvious implication is that $p_2=p_3=p_5$. The unbiased, unrestricted maximum likelihood estimates of the p_i's are $\hat{p}_1=.425$, $\hat{p}_2=.106$, $\hat{p}_3=.101$, $\hat{p}_4=.244$, and $\hat{p}_5=.124$.

Let p be the common value of p_2, p_3 and p_5 under the assumption the model holds. Then the maximum likelihood estimate of p is just $(p_2+p_3+p_5)/3=.110$ (Zehna, 1966). The maximum likelihood estimates of p_1 and p_4 are still .425 and .244 respectively. A chi square goodness-of-fit test yielded $X^2=1.055$ with one degree of freedom, and this is not significant at the .25 level. Thus, the model is reasonably consistent with the observed scores on item 7, and the maximum likelihood estimates of ζ, ζ_1 and ζ_2 are $\hat{\zeta}=.312$, $\hat{\zeta}_1=.55$, and $\hat{\zeta}_2=.134$, respectively.

The misinformation model just described assumes that examinees who incorrectly eliminate response B will choose the correct response on their fourth attempt. However, a slightly more general model can be applied. In particular, let γ be the probability that examinees with misinformation will choose the correct response on their fourth attempt once they learn that responses C, D, and E are incorrect. Then equations (3.5) and (3.6) become

$$p_4 = \gamma\zeta_2 + \zeta_1/5$$

$$p_5 = (1-\gamma)\zeta_2 + \zeta_1/5$$

Using equations (3.3) and (3.4) to estimate ζ_1, we now have that $\hat{\zeta}_1 = 5(.106 +.101)/2 = .5175$. Substituting this result in the remaining equations yields $\hat{\zeta} = .3215$, $\hat{\zeta}_2 = .161$, and $\hat{\gamma} = .873$.

4. AN EMPIRICAL CHECK OF WILCOX'S STRONG TRUE SCORE MODEL

Wilcox (1982) proposed a strong true-score model for answer-until-correct tests that can be described as follows: Consider a specific examinee responding to n items. Let y_i $(i=1,\ldots,t)$ be the number of items for which the examinee chooses the correct response on the ith attempt. Assume that the probability function of the y_i's is multinomial, i.e.,

$$f(y_1,\ldots,y_t|\theta_1,\ldots,\theta_t) = n! \prod_{i=1}^{t} \theta_i^{y_i}/y_i!$$

where the θ_i's are unknown parameters, $\Sigma\theta_i=1$, and $\Sigma y_i=n$. Wilcox assumes that for the population of examinees, the marginal distribution of y_1 is beta-binomial given by

$$f(y_1) = \binom{n}{y_1} \frac{B(r+y_1,\ n+s-y_1)}{B(r,s)} \tag{4.1}$$

where $r>0$ and $s>0$ are unknown parameters, and B is the beta function. Note that this assumption has proven to be useful when addressing various measurement problems (Wilcox, 1981b).

Next let $\xi=\theta_2/(1-\theta_1)$. Wilcox assumes that examinees with high ability are more likely to guess the correct response when they do not know. This assumption was expressed in terms of ξ by assuming that for the population of examinees, it is an increasing function of θ_1. In particular, $E(\xi|\theta_1)$ is assumed to be given by

$$c\int_0^1 \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1)\Gamma(\nu_2)} \theta^{\nu_1-1} (1-\theta)^{\nu_2-1} d\theta + (t-1)^{-1}$$

where c, ν_1 and ν_2 are unknown parameters satisfying $0\le c\le 1-(t-1)^{-1}$, $\nu_1>0$ and $\nu_2>0$. Since for a specific examinee

$$E(y_2|y_1, \theta_1, \theta_2) = \xi,$$

it follows that

$$E_\theta(y_2|y_1) = E(\xi|y_1) \tag{4.2}$$

where E_θ means expectation over the population of examinees (i.e., over the joint distribution of θ_1 and θ_2). This last result leads to an estimate of c, ν_1 and ν_2, and the details are given by Wilcox (1982).

First we tried fitting Wilcox's model to the items having t=5 distractors. As already pointed out, one of these items appears not to satisfy equation 2.1, and so it was eliminated. For the remaining 14 items, the parameters in equation 4.1 were estimated to be $\hat{r}=6.565$, and $\hat{s}=6.487$. The observed and expected frequencies are shown in columns two and three of Table 2. As can be seen, there is close agreement among the corresponding values, and a chi-square goodness-of-fit test is highly nonsignificant. Note that the items with t=3 alternatives could have been included, but they were analyzed separately in order to illustrate a simplification of the model that might be useful in certain situations.

Next, c, ν_1 and ν_2 were estimated to be $\hat{c}=.5$, $\hat{\nu}_1=1.2396$ and $\hat{\nu}_2=.5692$. The model assumes that for every examinee $\theta_2=(1-\theta_1)\xi$ where ξ is given by equation 4.2. This implies that the marginal distribution of y_2 is

$$f(y_2) = \int_0^1 f(y_2|\theta_1)g(\theta_1)d\theta_1, \tag{4.3}$$

where

$$f(y_2|\theta_1) = \frac{n!}{y_2!(n-y_2)!}\left((1-\theta_1)\xi\right)^{y_2}\left(1-(1-\theta_1)\xi\right)^{n-y_2} \tag{4.4}$$

and where from previous results, $g(\theta_1)$ is assumed to be a beta distribution with parameters $\hat{r}=6.565$ and $\hat{s}=6.487$. Thus, a check of the model is obtained by determining whether the right-hand side of equation 4.3 gives a good approximation to the observed marginal distribution of y_2. Equation 4.3 was evaluated with IBM (1971) subroutine DQG32. The observed and expected values for y_2 are shown in Table 2. As can be seen, equation 4.3 gives a reasonably good approximation to the observed frequencies, and a chi-square test is not significant at the.05 level.

A Random Guessing Model

It is interesting to see what happens when a random guessing model is assumed to hold. The expected frequencies for y_2 were computed, and they are shown in Table 2. It is clear that a random guessing model gives totally unsatisfactory results, and a goodness-of-fit test is highly significant. This result is consistent with results in Wilcox (1982) as well as Bliss (1980) and Cross & Frary (1977).

Analysis of Items with t=3 Alternatives

The analysis of the items with t=3 alternatives reveals that in some instances, a simpler version of Wilcox's model might be used. The motivation for this modification arose as follows: When estimating c, ν_1, and ν_2, the value of ξ is estimated at each of the y_1 values, and it is assumed that these values are strictly increasing. For the items having t=3 alternatives, the estimates of ξ corresponding to $y_1=2(1)15$ were .578, .577, .654, .615, .582, .564, .448, .574, .52, .636, .595, .552, and .57. There were no cases for $y_1=0$ or 1. If the estimation procedure used by

Wilcox is applied to these values, the results indicate a slight increase in ξ with increasing values of y_1, but the increase would seem to be too small to be concerned about. This suggests that a simpler model be considered where the ξ values are replaced by their average which is $\xi=.547$. Thus, for a specific examinee it is assumed that $\theta_2=.547(1-\theta_1)$. Next replace $(1-\theta_1)\xi$ with $.547(1-\theta_1)$ in equation 4.4, and replace $f(y_2|\theta_1)$ in equation 4.3 with the resulting expression. Again $g(\theta_1)$ was assumed to be a beta distribution, and the estimate of the parameters was found to be $\hat{r}=5.9877$ and $\hat{s}=4.5207$. The last two columns of Table 2 show the observed and expected frequencies of y_2, and the level of significance is greater than .1.

CONCLUDING REMARKS

Empirical investigations (Bliss, 1980; Cross & Frary, 1977) have shown that a random guessing model may be untenable, and it has been argued that such an assumption will frequently be unrealistic (e.g., Lord & Novick, 1968, p. 309). All indications are that guessing will be higher than random, and the strong true-score model described here is consistent with these results. Moreover, our common sense notion is that guessing should not be ignored, and in certain situations analytic results show that guessing can be a serious problem (van den Brink & Koele, 1980; Wilcox, 1980). Since all indications are that the assumptions about how examinees behave under answer-until-correct tests will frequently be consistent with observed test scores, perhaps it is now possible to deal with guessing in a more effective manner.

Another important point made by a referee is that investigators might want to collect pretest data under an AUC procedure even if the procedure is not to be used in operational versions of the test. Various possibilities are discussed elsewhere (Wilcox, 1981a, 1981c, in press a). These include the ability of estimating test item accuracy under conventional scoring procedures, and estimating the effectiveness of the distractors. If these values are judged to be too small, it might be possible to correct the problem by modifying or replacing some of the distractors.

Another situation where AUC tests might be useful involves the biserial correlation. When estimating this value, improved information about ζ might be useful (Ashler, 1979).

A third possible application is the empirical derivation of a formula score that corrects for guessing without assuming guessing is at random (Wilcox, 1982). Once certain parameters are estimated, this scoring formula can be used when the only available information is an examinee's observed number-correct score.

Finally, it is not being suggested that Wilcox's model be routinely applied. Instead, it is being argued that if the underlying assumptions seem reasonable, and if the observed test scores are consistent with these assumptions, then Wilcox's model might be considered when scoring and analyzing a test.

REFERENCES

Ashler, D. Biserial estimators in the presence of guessing. Journal of Educational Statistics, 1979, 4, 325-356.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Duncan, G. T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

IBM Application Program, System 1360. Scientific subroutines package (360-CM-03X) Version III programmer's manual. White Plains, NY: IBM Corporation Technical Publications Department, 1971.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Morrison, D. G., & Brockway, G. A modified beta-binomial model with applications to multiple choice and taste tests. Psychometrika, 1979, 44, 427-442.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

Wilcox, R.R. Determining the length of a criterion-referenced test. *Applied Psychological Measurement*, 1980, *4*, 425-446.

Wilcox, R.R. Solving measurement problems with an answer-until-correct scoring procedure. *Applied Psychological Measurement*, 1981, *5* to appear. (a)

Wilcox, R.R. A review of the beta-binomial model and its extensions. *Journal of Educational Statistics*, 1981, *6*, 3-32. (b)

Wilcox, R.R. *A polarization test for making inferences about the entropy of multiple-choice test items*. Unpublished technical report, Center for the Study of Evaluation, UCLA, 1981. (c)

Wilcox, R.R. Some empirical and theoretical results on an answer-until-correct scoring procedure. *British Journal of Mathematical and Statistical Psychology*, 1982, to appear.

Wilcox, R.R. Using results on k our of n system reliability to study and characterize tests. *Educational and Psychological Measurement*, in press. (a)

Wilcox, R.R. Determining the length of multiple-choice criterion-referenced tests when an answer-until-correct scoring procedure is used. *Educational and Psychological Measurement*, in press. (b)

Wilcox, R.R. A closed sequential procedure for answer-until-correct tests. *Journal of Experimental Education*, in press. (c)

TABLE 1

Number of Examinees Needing i (i=1,...,t) Attempts to Get the Correct Response

	ATTEMPTS				
ITEM	1	2	3	4	5
1	332	38	10	3	3
2	109	94	60	74	49
3	233	69	47	24	13
4	172	88	62	40	23
5	184	81	47	45	29
6	250	80	31	14	11
7	164	41	39	94	48
8	195	88	34	32	37
9	174	69	55	43	45
10	146	70	85	58	26
11	290	34	25	16	21
12	203	76	50	30	27
13	135	106	64	42	39
14	231	70	34	28	23
15	72	96	67	78	73
16	245	90	49		
17	168	125	91		
18	272	73	38		
19	228	140	14		
20	272	54	56		
21	220	89	72		
22	257	85	37		
23	308	47	22		
24	151	111	83		
25	121	130	119		
26	241	88	38		
27	235	79	50		
28	232	76	54		
29	101	121	140		
30	94	101	168		

TABLE 2

Observed an Expected Frequency

Value	Observed y_2 Frequencies $t=5$	Expected y_2 Frequencies $t=5$	Expected y_2 Frequencies Under Random Guessing	Observed y_2 Frequencies $t=3$	Expected y_2 Frequencies $t=3$, $\xi=.547$
0	23	24.51	68.16	16	13.24
1	64	70.77	115.82	33	35.25
2	94	99.78	101.73	54	54.24
3	82	90.04	60.09	70	62.97
4	72	57.65	26.29	45	58.97
5	31	27.54	8.91	48	46.57
6	12	10.09	2.39	36	31.45
7	5	2.87	.51	17	18.28
8	1	.63	.09	8	9.07
9				7	3.83
10				2	1.34

USING RESULTS ON k OUT OF n SYSTEM RELIABILITY TO STUDY AND CHARACTERIZE TESTS

Rand R. Wilcox

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles

ABSTRACT

For a specific achievement test item and a randomly selected examinee, let p be the probability of correctly determining whether the examinee knows the correct response. Various techniques have been proposed for estimating p. The purpose of this brief note is to describe and illustrate how results in the engineering literature on "k out of n system reliability" can be used to study and characterize tests based on the estimated values of p. In particular, we can empirically determine the minimum number of distractors required for multiple-choice tests. If we estimate p with an answer-until-correct scoring procedure, we can also determine the minimum number of examinees needed to be reasonably certain about whether γ is less than or greater than some predetermined constant, where $\gamma = \Sigma p_i$ and p_i is the value of p for the $i^{\underline{th}}$ item on an n-item test. In otherwords, we can determine whether the expected number of correct decisions on an n-item test is reasonably large.

Suppose we have a multiple-choice achievement test item that represents a particular skill. If an examinee chooses the correct response, we decide he/she has acquired the skill. As indicated in Section 2 of this paper, there are several methods for estimating the probability that for a typical examinee, we correctly decide whether the skill has been acquired. Usually however, these techniques have not been used to analyze tests that measure n skills, and they have not been used to *empirically* determine how many distractors we need for an item. The purpose of this paper is to illustrate how results in the engineering literature on "system reliability" can be used to help solve these problems. Section 3 reviews the results we will need. Included is a slight extension of an existing theorem which, as will be illustrated, is useful when addressing certain measurement problems. Section 4 describes six examples of how these techniques might be applied.

2. Methods for Estimating Item Accuracy

Under normal testing procedures it is impossible to estimate the probability of making an incorrect decision about whether an examinee has acquired a skill. In particular, there is no estimate of the probability of guessing the correct response when an examinee does not know, nor is there an estimate of the probability of knowing and being incorrect because of carelessness or a momentary distraction. However, there are circumstances under which these probabilities can be estimated.

One approach is suggested by Wilcox (1980). Consider a multiple-choice test item with t alternatives, one of which is correct. For a population of examinees, let ζ be the proportion who know the correct response, and let $\zeta_i (i=0,1,\ldots,t-2)$ be the proportion of examinees who do not know but who can eliminate i distractors. Suppose an answer-until-correct scoring procedure is used which means that examinees choose alternatives until the correct one is identified. If examinees who know are always correct on the first choice, and if examinees who do not know guess at random from among those distractors they cannot eliminate, then for a randomly selected examinee, the probability of a correct on the first alternative chosen is

$$\zeta + \sum_{j=0}^{t-2} \zeta_j/(t-j).$$

The probability of a correct on the i^{th} alternative chosen is

$$\tau_i = \sum_{j=0}^{t-i} \zeta_j/(t-j). \qquad (i = 2,\ldots, t).$$

Suppose we decide that a testee knows the answer if the first alternative chosen is correct. The probabilities of the four possible outcomes are shown in Table 1.

TABLE 1

Four Possible Outcomes of a Randomly Selected Examinee Responding to an Item

		Decision	
		Knows	Does Not Know
Latent	Knows	ζ	0
State	Does Not Know	τ_2	$\sum_{i=2}^{t} \tau_i$

Thus, the probability of a correct decision for a randomly selected examinee is

$$p = \zeta + \sum_{i=2}^{t} \tau_i$$

It can be shown that for fixed ζ, p attains its maximum value when guessing is at random, i.e., $\tau_0 = 1-\zeta$.

For a random sample of N examinees, let z_i be the number of times a correct response is given on the $i^{\underline{th}}$ try. Then $(z_1-z_2)/N$, and $(n-z_2)/N$ are unbiased maximum likelihood estimate of ζ and p respectively. Unbiased maximum likelihood estimates of the ζ_i's are also readily obtained as is illustrated by Wilcox (1980) for the special case $t=4$.

Another way to estimate the accuracy of decisions about whether the typical examinee has acquired a particular skill is with latent structure models. Macready and Dayton (1977) illustrate this for the case of equivalent items. Two items are defined to be equivalent if a randomly sampled examinee knows both or neither one. In addition to including guessing, the model used by Macready and Dayton allows for the event of an examinee knowing and being incorrect.

There are four methods for checking the assumption of equivalent items (Macready and Dayton, 1977; Hartke, 1978; Baker and Hubert, 1977; and Wilcox, in press, a). If the assumption of equivalent items is contraindicated by the data, we might still use a latent structure model, but one based on less stringent assumptions. In particular, we might assume items are hierarchically related which contains the assumption of equivalent items as a special case (Wilcox, in press, b). Dayton and Macready (1976) describe a general approach to hierarchically related items.

3. Review of "Reliability" Theory

Suppose a test measures n skills. For a randomly selected examinee let $x_i=1$ if a correct decision in made about whether the $i^{\underline{th}}$ skill has been acquired; otherwise, $x_i=0$. Also, let $p_i=Ex_i$ where the expectation is taken over the population of examinees. As noted in the previous section, p_i can be estimated under various circumstances. We define the k out of n reliability of a test, ρ_k, to be the probability of making at least k correct decisions for a randomly selected examinee. Symbolically, $\rho_k=Pr(\Sigma x_i \geq k)$.

Some readers might object to defining test reliability in the manner described above since it differs from the usual definition of reliability in classical test theory. The reason we do so is because it is consistent with the usual definition of system reliability that is applied to engineering problems (e.g., Barlow and Proschan, 1975; Marshall and Olkin, 1979, p. 402).

The purpose of this section is to list some results about ρ_k. Except for Theorem 6, these results are not new, but they are not typically applied to measurement problems, and so we describe them here for the convenience of the reader.

First we note that if x_i is independent of x_j, $i \neq j$,

$$\rho_k = \sum_{\underline{x}:S \geq k} \prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i} \qquad (1)$$

where $\underline{x}=(x_1,\ldots,x_n)$ and $S=\Sigma x_i$. For some cases, (1) is easily computed, for example, when n is small or k=n, but frequently ρ_k is difficult to calculate. Another and perhaps more serious problem is that the x_i's might not be independent. In this case, determining ρ_k is more difficult. For these reasons, efforts have been made to find ways to approximate ρ_k, and to determine its properties.

Theorem 1. ρ_k is strictly increasing in each p_i. A proof is given by Barlow and Proschan (1975, p. 22).

Theorem 2. If $\mathrm{cov}(x_i, x_j) \geq 0$, $i \neq j$, then

$$\prod_{i=1}^{n} p_i \leq \rho_k \leq 1 - \prod_{i=1}^{n} (1-p_i)$$

This is a special case of a result given by Barlow and Proschan (1975, p. 34).

Theorem 3. If $\mathrm{cov}(x_i, x_j) \geq 0$,

$$\max_{\underline{x}:S=k} \prod_{i=1} p_i^{x_i} \leq \rho_k$$

This follows from Theorem 3.9 in Barlow and Proschan (1975, p. 37).

Definition: For any vector $\underline{a}$, let $a_{(1)} \geq a_{(2)} \geq \cdots \geq a_{(n)}$ be the elements of $\underline{a}$ written in descending order. The vector $\underline{a}$ is said to be majorized by the vector $\underline{b}$ (b majorizes $\underline{a}$) if

$$\sum_{i=1}^{k} a_{(i)} \leq \sum_{i=1}^{k} b_{(i)} \quad , \quad i = 1, \ldots, n-1$$

and

$$\sum_{i=1}^{n} a_{(i)} = \sum_{i=1}^{n} b_{(i)} \tag{2}$$

Symbolically, $\underline{b}$ majorizes $\underline{a}$ is written $\underline{a}\leq^{m}\underline{b}$. If (2) is replaced by

$$\sum_{i=1}^{n} a_{(i)} \leq \sum_{i=1}^{n} b_{(i)} \quad ,$$

$\underline{b}$ weakly majorizes $\underline{a}$, which we write $\underline{a}\leq^{wm}\underline{b}$.

Theorem 4. Suppose we have two test forms where p_i is defined as before and p_i^* is the corresponding probability on the other test. Suppose x_i is independent of x_j, $i\neq j$, for both test forms. Let $r_i=-\log p_i$, and $r_i^*=-\log p_i^*$, where log is the natural logarithm. If $\underline{r}^*\leq^{m}\underline{r}$, then

$$\rho_k(\underline{p}^*) \leq \rho_k(\underline{p})$$

with equality holding when k=n. A proof is given by Pledger and Proschan (1971).

A corollary of Theorem 4 is that $\rho_k(\underline{p}) \geq \rho_k(p_G,\dots p_G) = \sum_{x=k}^{n}\binom{n}{x}p_G^{\,x}(1-p_G)^{n-x}$ for any $\underline{p}$, where $p_G = \left[\prod_{i=1}^{n} p_i\right]^{\frac{1}{n}}$ is the geometric mean of the p_i's.

Theorem 5. Pledger and Proschan (1971) also show that if $R_i=(1-p_i)/p_i$, and $R_i^* = (1-p_i^*)/p_i^*$, then $\underline{R}^*\leq^{m}\underline{R}$ implies that

$$\rho_k(\underline{p}^*) \leq \rho_k(\underline{p})$$

We should remark that $\underline{r}\leq^{m}\underline{r}^*$ does not imply that $\underline{R}\leq^{m}\underline{R}^*$, nor is the converse true.

Theorem 6. Let r, r^*, R and R^* be defined as in theorems 4 and 5. Suppose x_i is independent of x_j, $i \neq j$, for both test forms, that $\underline{r}$ does not majorize $\underline{r}^*$, but that for some $\underline{c}$, $\underline{r} \geq^m \underline{r}' = \underline{r}^* - \underline{c}$, where $c_i \geq 0 (i=1,\ldots,n)$. Then $\rho_k(\underline{p}) \geq \rho_k(\underline{p}^*)$ for any k. The same is true if $\underline{R} \geq^m \underline{R}' = \underline{R}^* - \underline{c}$ for some $\underline{c}$, $c_i \geq 0$.

Proof: Theorem 4 says that $\underline{r} \geq^m \underline{r}'$ implies that $\rho_k(\underline{p}) \geq \rho_k(\underline{p}')$. Also, $\underline{r}' = \underline{r}^* - \underline{c}$ means that $p_i' = p_i^* + b_i$ for some appropriately chosen $b_i \geq 0$, where p_i' is the p_i value corresponding to r_i'. Thus, by Theorem 1, $\rho_k(\underline{p}') \geq \rho_k(\underline{p}^*)$. The proof is exactly the same for $\underline{R}$ and $\underline{R}^*$.

Theorem 7. If x_i is independent of x_j, $i \neq j$, Hoeffding (1956) shows that

$$\rho_k(\underline{p}) \geq \sum_{j=k}^{n} \binom{n}{j} \bar{p}^j (1-\bar{p})^{n-j} \quad , \quad k \leq n\bar{p}$$

and

$$\rho_k(\underline{p}) \leq \sum_{j=k}^{n} \binom{n}{j} \bar{p}^j (1-\bar{p})^{n-j} \quad , \quad k-1 \geq n\bar{p}$$

where $\bar{p} = n^{-1}\Sigma p_i$. (See, also, Gleser, 1975.)

4. Applications

As previously indicated, the purpose of this paper is to illustrate how the above theorems can be applied to certain measurement problems. This we now do.

Example 1: Suppose multiple-choice test items are being used, and that we want

$$\rho_k \geq P^* \tag{3}$$

for some positive $P^*<1$. What is the minimum number of distractors required.

To solve this problem, suppose guessing is at random, and that an examinee behaves as assumed under the answer-until-correct scoring procedure described in Section 2. As was pointed out, for fixed ζ, p_i is maximized when guessing is at random. Furthermore, the value of p_i is an increasing function of $\zeta^{(i)}$, the proportion of examinees who know the answer to the ith item. Since for any i, $\zeta^{(i)}$ is unknown, first consider the value of $\zeta^{(i)}$ that minimizes p_i. This is $\zeta^{(i)} = 0 (i=1,\ldots,n)$. Suppose x_i is independent of x_j, $i \neq j$. If the same number of distractors is used for each item, then $p_1=p_2=\ldots=p_n=p$, say, and

$$\rho_k = \sum_{x=k}^{n} \binom{n}{x} p^x (1-p)^{n-x} \tag{4}$$

Let p_0 be the value of p for which (4) equals P^*. Then the number of required distractors is $t = (1-p_0)^{-1}$ since, when $\zeta=0$, $p=1-t^{-1}$. For example, if $P^* = .93$, $n=10$, and $k=8$, then $p_0 = .9$ and $t=10$. Of course, in practice, this is an extremely large number of distractors. However, $\zeta^{(i)}=0$ $(i=1,\ldots,n)$ is highly unlikely, and so in reality, a smaller number of distractors would be needed when guessing is at random.

To illustrate Theorem 3, consider the more general case where $\mathrm{cov}(x_i, x_j) \geq 0$. If again $p_1 = \ldots = p_n = p$, to guarantee (4) we determine p such that $p^k = P^*$. From Theorem 3 it follows that the required number of distractors is $(1-p)^{-1}$. If, for example, $P^* = .93$, $n=10$ and $k=8$, then $(.991)^8 = .93$, and so $t=111$. To reiterate, this value of t is based on an unrealistic value for the $\zeta^{(i)}$'s. More realistic situations are considered below. Our goal here is to illustrate Theorem 3 in a simple manner.

Example 2: We consider the same situation as in example 1, but we assume information about the ζ's is available. More specifically, suppose the ζ's have been estimated to be $\zeta^{(1)}=.5$, $\zeta^{(2)}=.6$, $\zeta^{(3)}=\zeta^{(4)}=.75$, $\zeta^{(5)}=\zeta^{(6)}=\zeta^{(7)}=.85$, $\zeta^{(8)}=.9$ and $\zeta^{(9)}=\zeta^{(10)}=.95$. To determine the minimum number of distractors, again assume guessing is at random, that $\mathrm{cov}(x_i, x_\zeta) \geq 0$, and $k=8$. To simplify the illustration, suppose the same number of distractors is to be used for each item. Since p_i is an increasing function of ζ, and since when guessing is random $p_i = \zeta + (1-\zeta)(1-t^{-1})$, we have, by an application of Theorem 3, that a lower bound to ρ_8 is

$$(.75+.25(1-t^{-1}))^2(.85+1.5(1-t^{-1}))^3(.9+.1(1-t^{-1}))(.95+.05(1-t^{-1}))^2 \quad (5)$$

Thus, we can guarantee $\rho_8 \geq P^*$ by finding the smallest t such that (5) is greater than or equal to P^*. Table 2 gives the value of (5) for $t=4(1)8$.

TABLE 2

Values of (5) for $t=4(1)8$ and $k=6$, 7 and 8

		t				
		4	5	6	7	8
k	8	.745	.79	.82	.85	.87
	7	.79	.83	.85	.88	.89
	6	.85	.88	.90	.91	.92

These results are more encouraging than those in example 1, but having more than 4 or 5 equally attractive distractors promises to be difficult in practice.

We note that the lower bound to ρ_k in Theorem 3 can be very sensitive to the value of k. Table 3 also gives the value of (5) for t=4(1)8 for both k=7 and k=6.

Next suppose that x_i and x_j are independent, $i \neq j$. From the corollary to Theorem 4, a lower bound to ρ_8 is

$$\sum_{x=8}^{10} \binom{10}{x} p_G^x (1 - p_G)^{10-x} \qquad (6)$$

Since we are still assuming guessing is at random, $p_i = \zeta + (1-\zeta)(1-t^{-1})$ and equation (6) is easily calculated for any t. The values of p_G corresponding to t=2,3,4 are respectively, .894, .942 and .948. Substituting these values in (6), it follows that for t=2, $\rho_8 \geq .915$, for t=3, $\rho_8 \geq .983$, and for t=4, $\rho_8 \geq .987$.

If instead we apply Theorem 7, the values of $\overline{p}$ corresponding to $t=2,3$ and 4 are .8975, .9317, and .9488, and the resulting lower bounds to ρ_8 are .925, .973, and .988, respectively. As is evident, the lower bound to ρ_8 for the case $t=2$ is higher than it was using the corollary to Theorem 4, but for $t=3$ and 4, the lower bounds are about the same.

In contrast to the previous illustrations, test accuracy is very high using a "normal" number of distractors. An interesting feature of the illustration just given is that there seems to be little reason for using $t=4$ distractors, rather than $t=3$, since the increase in ρ_8 is minimal at best. Note, however, that t was derived under the assumption of random guessing. If examinees have partial information, the p_i values will be lower which in turn will lower the value of ρ_8. As mentioned in section 2, an answer-until-correct scoring procedure can be used to check for partial information, and to estimate the p_i's.

Example 3: The situation is assumed to be the same as in example 2, except that we want to allow for the possibility of having a different number of distractors across items. Assuming x_i is independent of x_j, $i \neq j$, the simplest approach to guaranteeing $\rho_8 \geq P^*$ is to determine the smallest t for each item such that $p_i \geq (p_0)^{1/n}$ where p_0 is the value of p_G in the corollary to Theorem 4 such that $\sum_{x=8}^{n} \binom{n}{x} p_G^{*}(1-p_G)^{n-x} = P^*$. If, for example, $P^*=.95$, $p_0=.915$. It follows that for $\zeta=.5$, .6, .75, .85, .9 and .95, the corresponding values of t are 6,5,5,3,2,2, respectively. (We assume that a minimum of $t=2$ distractors are used.)

For $t=3$ in example 2, assuming x_i is independent of x_j, $i \neq j$, and that guessing is at random, $\underline{p} = (.834, .871, .92, .92, .952, .952, .968, .984, .984)$

implying that $p_G=.942$ and so $\rho_8\geq.983$. In example 3, using the indicated t values, p is given by $p'=(.917, .92, .95, .95, .952, .952, .952, .95, .975, .975)$, $p_G=.959$, and so $\rho_8\geq.993$. This suggests that the k out of n reliability with the latter test form is higher than the first--but this has not been established. The corollary to Theorem 4 gives a lower bound to ρ_k, but it has not been shown that the lower bound indicates which test form is more accurate. If it had been true that $p_i'\geq p_i$ $(i=1,...,10)$, the test form in example 3 would be more accurate according to Theorem 1, but it is evident that this is not the case. However, by applying Theorem 6, it can be shown that the latter test form has a higher value for ρ_8.

Example 4: As mentioned in section 2, Macready and Dayton (1977) examine a latent structure model that can be used to estimate p. Included in their discussion is a solution to the following problem: When measuring a particular skill, how many items are needed, and what passing should we use, so that the probability of making a correct decision about whether a typical examinee has acquired a particular skill, i.e., the value of p, is reasonably close to one.

As before, let ζ be the proportion of examinees who have acquired the skill, and for a randomly selected examinee, let α=Pr (in correct response| examinee knows) and let β=Pr(correct response|examinee does not know). Suppose n equivalent items are to be used to measure a skill. Macready and Dayton provide a table of n values and passing scores corresponding to various values of ζ, α, and β. For example, if $\zeta=.6$, $\alpha=.05$ and $\beta=.3$, and if we want, with probability at least .95, to correctly determine whether a randomly selected examinee has acquired the skill being measured, Table I in Macready and Dayton (1977) says we need to use n=4 items with a passing score of 3.

Using the results in section 3, we can extend the technique proposed by Macready and Dayton to tests that measure m skills. As a simple illustration, suppose we have 4 skills, and the number of items (and passing score) corresponding to these four skills are 4(3), 5(4), 4(3) and 7(5), respectively. For the first skill, for example, we have four items, and if an examinee gets at least 3 correct, we decide he/she has acquired the skill. Further, suppose that the estimation procedures described by Macready and Dayton are applied, and the four estimates of ζ are .4, .5, .6, and .75; the corresponding estimates of α are .05, .1, .05, and .1; and the estimates of β are .2, .3, .4, and .4. The minimum probability of a correct decision associated with the four skills can be read from Macready and Dayton's Table I (assuming a loss ratio of one), and they are .95, .9, .9, .95. Making the appropriate independence assumption, Theorem 7 says that, for a randomly selected examinee, a lower bound to the probability of making at least 3 correct decisions for the four skills is .97.

Example 5: A proficiency test is designed to measure m skills. For each skill, a decision is made about whether an examinee knows the correct response. How many items per skill do we need so that for the m skills, at least k correct decisions are made for the typical examinee. This problem is similar to previous illustrations. It can be solved using the results in section 3 in conjunction with the techniques described by Macready and Dayton.

Example 6: Suppose every examinee behaves as described in section 2 under the answer-until-correct scoring procedure, and that x_i is independent of x_j, $i \neq j$. Let $\gamma = \Sigma E x_i = \Sigma p_i$ be the expected number of correct decisions on an n-item test, i.e., the number of times we expect to correctly determine whether a typical examinee knows the answer to an item. We consider the problem of determining whether γ is reasonably large, say greater than or equal γ_0, a known constant. For a random sample of N examinees, let $w_{ij}=0$ if the jth examinee is correct on the second attempt of the ith item; otherwise $w_{ij}=1$. From section 2, $Ew_{ij}=p_i$, $j=1,\ldots,N$, and so we decide $\gamma \geq \gamma_0$ if $\hat{\gamma} = \sum_i \sum_j w_{ij}/N \geq \gamma_0$; otherwise we decide $\gamma < \gamma_0$. How large must N be so that we can be reasonably certain of making a correct decision about whether γ is greater than or less than γ_0?

Note that the situation is similar to one considered by Fhanér (1974) and Wilcox (1979). The main difference is that rather than a binomial model, here we have a compound binomial distribution.

Following Fhanér (1974) suppose we want to choose the smallest N so that when $\gamma \geq \gamma_0 + \delta^*$,

$$\Pr(\hat{\gamma} \geq \gamma_0) \geq T \tag{6}$$

and when $\gamma \leq \gamma_0 - \delta^*$,

$$\Pr(\hat{\gamma} < \gamma_0) \geq T \tag{7}$$

where $\frac{1}{2} < T < 1$. Fhaner assumes $\delta^* > 0$, but we require $\delta^* \geq 1$ so that we can apply Theorem 7. In particular, for $\gamma \geq \gamma_0 + \delta^*$, and for $N=1$,

$$\Pr(\hat{\gamma} \geq \gamma_0) = \cancel{p_{\lceil \gamma_0 \rceil}} \geq \sum_{x=\lceil \gamma_0 \rceil}^{n} \binom{n}{x} \left(\frac{\gamma_0 + \delta^*}{n} \right)^x \left(\frac{n - \gamma_0 - \delta^*}{n} \right)^{n-x} \tag{8}$$

where $\lceil \gamma_0 \rceil$ is the smallest integer $\geq \gamma_0$. Again applying Theorem 7 (the second inequality) we have that for $\gamma \leq \gamma_0 - \delta^*$.

$$\Pr(\hat{\gamma} < \gamma_0) = 1-\rho_{[\gamma_0]} \geq \sum_{x=0}^{[\gamma_0]-1} \binom{n}{x} \left(\frac{\gamma_0-\delta^*}{n}\right)^x \left(\frac{n-\gamma_0+\delta^*}{n}\right)^{n-x} \tag{9}$$

Let η_1 and η_2 be the right-hand side values of (8) and (9) respectively. From the above results, we have that for a random sample of N examinees, when $\gamma \geq \gamma_0+\delta^*$

$$\Pr(\hat{\gamma} > \gamma_0) \geq \sum_{y=[N\gamma_0]}^{N} \binom{N}{y} \eta_1^{\,y} (1-\eta_1)^{N-y} \tag{10}$$

and when $$\gamma \leq \gamma_0 - \delta^*$$

$$\Pr(\hat{\gamma} < \gamma_0) \geq \sum_{y=0}^{[N\gamma_0]-1} \binom{n}{y} \eta_2^{y} (1-\eta_2)^{N-y} \tag{11}$$

Thus, we can guarantee both (6) and (7), regardless of the actual value of γ, by choosing the smallest N so that the right-hand side of both (10) and (11) are greater than or equal to T.

As a more specific example, suppose we have an n=10 item test, that an answer-until-correct scoring procedure is to be used to estimate γ and we want to determine the minimum number of examinees we need in order to correctly determine whether γ is above or below $\gamma_0=7$. In particular, suppose $\delta^*=1$, and that if $\gamma \geq \gamma_0+\delta^*$ we want $\Pr(\hat{\gamma} \geq \gamma_0) \geq T=.9$, and if $\gamma \leq \gamma_0-\delta^*$ we want $\Pr(\hat{\gamma} < \gamma_0) \geq T$, regardless of the actual of γ.

From (8) and (9), $\eta_1 = \sum_{x=7}^{10} \binom{10}{x} .8^x .2^{n-x} = .897$ and $\eta_2 = \sum_{x=0}^{6} \binom{10}{x} .6^x .4^{n-9} = .618$. Substituting these values into (10) and (11), it can be verified that the minimum N required is 67.

References

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. Journal of Educational Statistics, 1977, 2, 217-233.

Barlow, R. E., & Proschan, F. Statistical theory of reliability and life testing: Probability models. New York: Holt, Rinehart & Winston, 1975.

Dayton, C. M., & Macready, G. B. A probabalistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Fhanér, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Gleser, L. J. On the distribution of the number of successes in independent trials. Annals of Probability, 1975, 3, 182-188.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 1978, 15, 43-47.

Hoeffding, M. On the distribution of the number of successes in independent trials. Annals of Mathematical Statistics, 1956, 27, 713-721.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marshall, A. W., & Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

Pledger, G., & Proschan, F. Comparisons of order statistics and of spacings from heterogeneous distributions. In J. S. Rustagi (Ed.) Optimizing Methods in Statistics. New York: Academic Press, 1971.

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Center for the Study of Evaluation, University of California, Los Angeles, 1980.

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement, 1979, 31, 13-22.

Wilcox, R. R. The single administration estimate of the proportion of agreement of a proficiency test scored with a latent structure model. Educational and Psychological Measurement, in press. (a)

Wilcox, R. R. Some results and comments on using latent structure models to measure achievement. Educational and Psychological Measurement, 1980, in press. (b)

BOUNDS ON THE K OUT OF N RELIABILITY OF A TEST, AND AN EXACT TEST FOR RANDOM GUESSING

Rand R. Wilcox

Department of Psychology
University of Southern California

and

The Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

Consider an n-item multiple choice test where it is decided that an examinee knows the answer if and only if he/she gives the correct response. The k out of n reliability of the test, ρ_k, is defined to be the probability that for a randomly sampled examinee, at least k correct decisions are made about whether the examinee knows the answer to an item. The paper describes and illustrates how an extension of a recently proposed latent structure model can be used in conjunction with results in Sathe et al. (1980) to estimate upper and lower bounds on ρ_k. A method of empirically checking the model is discussed. Included is an exact test of whether guessing is at random.

Consider a randomly sampled examinee responding to a multiple-choice test item. In mental test theory there are, of course, many procedures that might be used to analyze this item. One approach might be as follows. Suppose a conventional scoring procedure is used where it is decided that an examinee knows the correct response if the correct alternative is chosen, and that otherwise the examinee does not know. If it were possible to estimate the probability, τ, of correctly determining an examinee's latent state (whether he/she knows the correct response) based on the above decision rule, this would give an indication of how well the item is performing for the typical examinee. The obvious problem is that under normal circumstances, there is no way of estimating this probability unless additional assumptions are made. One approach is to assume that examinees guess at random among the alternatives when they do not know the answer. If this knowledge or random guessing model holds, τ is easily estimated. However, empirical investigations (Bliss, 1980; Cross & Frary, 1977) suggest that this assumption will frequently be violated, and some related empirical results (Wilcox, 1982, in press a) indicate that such a model can be entirely unsatisfactory for other reasons as well.

Another approach is to use a latent structure model, and many such models have been proposed for measuring achievement (e.g., Brownless & Keats, 1958; Marks & Noll, 1967; Knapp, 1977; Dayton & Macready, 1977, 1980; Macready & Dayton, 1977; Wilcox, 1977a, 1977b, 1981a; Bergan et al., 1980). The choice of a model depends on what one is willing to assume in a particular situation. These models make it possible to estimate errors at the item level such as

$$\beta = \Pr(\text{randomly selected examinee gives the correct response} \mid \text{examinee does not know}) \qquad [1]$$

which in turn yields an estimate of τ. An illustration is given in a later section. (For a review of latent structure models vis-à-vis criterion-referenced tests, see Macready and Dayton, 1981.) For some recent general comments on using latent structure models to measure achievement, see Molenaar (1981) and Wilcox (1981b).

Assume for the moment that for each item on an n-item test, an estimate of τ can be made. Let $x_i = 1$ if a correct decision is made on the ith item for a randomly selected examinee; otherwise $x_i = 0$. Then $E(x_i) = \tau_i$ $(i = 1, \ldots, n)$ is the probability of a correct decision on the ith item where the expectation is taken over the population of examinees.

Within the framework just described, how should an n-item test be characterized? An obvious approach is to use

$$\mu = E(\Sigma x_i) = \tau_i \qquad [2]$$

which is the expected number of correct decisions among the n items.

Knowing μ might not be important for certain types of tests, but surely it is important for some achievement tests. However, even if μ is known exactly, it would be helpful to have some additional related information about Σx_i. For instance, a test constructor would have a better idea of how the test performs if $VAR(\Sigma x_i)$ could be determined. The problem is that $VAR(\Sigma x_i)$ depends on $COV(x_i, x_j)$, but this last quantity is not known, and at present there is no way of estimating it. An alternative approach is to use the k out of n reliability of the test (Wilcox, in press b) which is given by

$$\rho_k = \Pr(\Sigma x_i \geq k) \ . \qquad [3]$$

In other words, if the goal of a test is to determine which of n items an examinee knows, and if a conventional scoring procedure is used, ρ_k is the probability of making at least k correct decisions for the typical examinee.

Suppose, for example, $n = 10$ and μ is estimated to be 7. Thus, the expected number of correct decisions is 7, but there is no information about the likelihood that at least 7 correct decisions will be made. If ρ_k were known, a test constructor would have some additional and useful information for judging the accuracy of the test. ρ_k might also be used as follows. Suppose it is desired to have $\rho_8 \geq .9$. If μ is estimated to be 9.1, this is encouraging, but it is not clear what implications this has in terms of making at least 8 correct decisions for the typical examinee.

It is not being suggested that determining ρ_k is important for every test that might be constructed, but certainly it is important in various situations. For example, when measuring progress through an instructional program, surely it is desirable to determine which of the skills represented by the items on the test have or have not been acquired by an examinee. An estimate of ρ_k yields information about how well a test performs this goal.

If x_i is independent of x_j, $i \neq j$, an exact expression for ρ_k is available via the compound binomial distribution. Perhaps there are situations where this independence might be assumed, but it is evident that this independence will not always hold. If it can be assumed that $\mathrm{COV}(x_i, x_j) \geq 0$, bounds on ρ_k are available (Wilcox, in press b). Recently Sathe, Pradhan, and Shah (1980) derived bounds on ρ_k that make no assumption about COV(x_i,x_j). The main point of this paper is that these bounds can be estimated using an extension of an answer-until-correct (AUC) scoring procedure proposed by Wilcox (1981a). The paper also indicates how an exact test can be made of certain implications of the new model. This procedure can also be used to make an exact test of whether guessing is at random. (For an asymptotic test, see Weitzman, 1970.) Finally, the paper includes some comments on how a test might be modified when μ or ρ_k is judged to be too small.

An Extension of an Answer-Until-Correct Scoring Procedure

As just indicated, an extension of results in Wilcox (1981a) is needed in order to apply the bounds derived by Sathe et al. (1980). First, however, it is helpful to briefly review the procedure and basic assumptions in Wilcox (1981a).

Consider a specific test item having t alternatives from which to choose, one of which is the correct response. Assume examinees respond according to an AUC scoring procedure. This means that examinees choose an alternative, and they are told immediately whether the correct response has been identified. If they are incorrect another response is chosen, and this process continues until they are successful. Special forms are generally available for administering AUC tests which make these tests easy to use in the classroom.

Let ζ_{t-1} be the proportion of examinees who know the correct response, and let ζ_i ($i = 0, \ldots, t-2$) be the proportion of examinees who can eliminate i distractors given that they do not know. Wilcox (1981a) assumes that examinees eliminate as many distractors as they can, and then choose at random from among those that remain. If p_i

is the probability of choosing the correct response on the ith attempt, then

$$p_i = \sum_{j=0}^{t-i} \zeta_j/(t - j) \quad (i=1, \ldots, t). \qquad [4]$$

Note that the model assumes that at least one effective distractor is being used. Put another way, no distinction is made between examinees who know the answer and examinees who can eliminate all of the distractors.

Assuming the model holds,

$$\zeta_{t-1} = p_1 - p_2 \qquad [5]$$

and

$$\tau = 1 - p_2 \, . \qquad [6]$$

If in a random sample of N examinees, y_i examinees are correct on their ith attempt, $\hat{p}_i = y_i/N$ is an unbiased estimate of p_i which yields an estimate of ζ_{t-1} and τ.

Although empirical studies suggest that this model will frequently be reasonable (Wilcox, 1982, in press a), there are instances where this will not be the case. For example, some items might require a misinformation model, and an appropriate modification of the AUC scoring procedure has been proposed (Wilcox, in press a). Further comments on this problem are made in a later section of the paper.

Consider any two items on an n-item test, say items i and j. Applying results in Sathe et al. requires an estimate of $\tau_{ij}=Pr(x_i=1,x_j=1)$, i.e., the joint probability of making a correct decision for both items i and j. The remainder of this section outlines how this might be done.

It is assumed that an examinee's guessing rate is independent over the items that he/she does not know. This means, for example, that if an examinee can eliminate all but 2 alternatives on item i, and all but

3 alternatives on item j, the probability of choosing the correct response on the first attempt of both items is $(1/2)(1/3) = 1/6$.

For the two items under consideration, let p_{km} $(k, m = 1, \ldots, t)$ be the probability that a randomly selected examinee chooses the correct response on the kth attempt of the first item, and the correct response on the mth attempt of the second. If ζ_{gh} is the proportion of examinees who can eliminate g distractors from the first item and h distractors from the second $(g, h = 1, \ldots, t-1)$, then

$$p_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \zeta_{ij}/[(t - i)(t - j)] \qquad [7]$$

The last expression can be used to express $\zeta_{t-1,t-1}$ in terms of the p_{km}'s which can be used to estimate $\zeta_{t-1,t-1}$. Note that if the first item has t' alternatives, $t' \neq t$, simply replace t-k with t'-k in equation 7.

To clarify matters, consider the special case $t = 3$. Equation 7 says that

$$p_{11} = \zeta_{22} + \zeta_{21}/2 + \zeta_{20}/3 + \zeta_{12}/2 + \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{02}/3 + \zeta_{01}/6 + \zeta_{00}/9 \qquad [8]$$

$$p_{12} = \zeta_{21}/2 + \zeta_{20}/3 + \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{01}/6 + \zeta_{00}/9 \qquad [9]$$

$$p_{13} = \zeta_{20}/3 + \zeta_{10}/6 + \zeta_{00}/9 \qquad [10]$$

$$p_{21} = \zeta_{12}/2 + \zeta_{02}/3 + \zeta_{11}/4 + \zeta_{01}/6 + \zeta_{10}/6 + \zeta_{00}/9 \qquad [11]$$

$$p_{22} = \zeta_{11}/4 + \zeta_{10}/6 + \zeta_{01}/6 + \zeta_{00}/9 \qquad [12]$$

$$p_{23} = \zeta_{10}/6 + \zeta_{00}/9 \qquad [13]$$

$$p_{31} = \zeta_{02}/3 + \zeta_{01}/6 + \zeta_{00}/9 \qquad [14]$$

$$p_{32} = \zeta_{01}/6 + \zeta_{00}/9 \qquad [15]$$

$$p_{33} = \zeta_{00}/9 \qquad [16]$$

Thus, starting with equation 16

$$\zeta_{10} = 9p_{33} \tag{17}$$

$$\zeta_{01} = 6(p_{32} - p_{33}) \tag{18}$$

and eventually ζ_{22} can be expressed in terms of the p_{km}'s. Replacing the p_{km}'s with their usual unbiased estimate yields an estimate of ζ_{22} say $\hat{\zeta}_{22}$. But it can be seen that for the two items under consideration (items i and j),

$$\tau_{ij} = \zeta_{22} + 1 - p_{11} \tag{19}$$

Replacing ζ_{22} and p_{11} with $\hat{\zeta}_{22}$ and $\hat{p}_{11}$ yields an estimate of $\tau_{ij} = \Pr(x_i=1, x_j=1)$, say $\hat{\tau}_{ij}$. For arbitrary t, τ_{ij} is given by equation 19 with ζ_{22} replaced with $\zeta_{t-1,t-1}$.

Bounds on ρ_k

This section illustrates how the results in the previous section can be used to estimate bounds on ρ_k. First, however, results in Sathe et al. (1980) are summarized.

Recall that $\mu = \Sigma\tau_i$ and let

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \tau_{ij} \tag{20}$$

$$U_k = \mu - k \tag{21}$$

and

$$V_k = (2S - k(k - 1))/2 \tag{22}$$

Then,

$$\rho_k \geq \frac{2V_{k-1} - (k - 2)U_{k-1}}{n(n - k + 1)} \tag{23}$$

If $2V_{k-1} < (n + k - 2)U_{k-1}$, then

$$\rho_k \geq \frac{2((k^* - 1)U_{k-1} - V_{k+1})}{(k^* - k)(k^* - k + 1)} \tag{24}$$

where $k^* + k - 3$ is the largest integer in $2V_{k-1}/U_{k-1}$. Two upper bounds on ρ_k are also given. The first is

$$\rho_k \leq 1 + ((n + k - 1)U_k - 2V_k)/kn \qquad [25]$$

and the second is that if $2V_k < (k - 1)U_k$,

$$1 - 2\,\frac{(k^* - 1)U_k - V_k}{(k - k^*)(k - k^* + 1)} \leq \rho_k \qquad [26]$$

where $k^* + k - 1$ is the largest integer in $2V_k/U_k$.

An Illustration

To illustrate how ρ_k might be applied and interpreted, observations of seven items were analyzed according to the procedure outlined above. Each item had two distractors, and they were found to be consistent with the assumptions of the answer-until-correct scoring model. (See Wilcox, 1981a). Table 1 shows the observed frequencies for the first two items. The question to be answered is if these seven items are taken to be the whole test, do they give reasonably accurate information about what the typical examinee knows?

Generally, when estimating ζ_{22} there is no need to estimate all of the ζ's in equations 8-16. For the situation at hand, ζ_{22} can be estimated as follows. First compute

$$\hat{\zeta}_{02}/3 = \hat{p}_{31} - \hat{p}_{32} \qquad [27]$$

for the data in Table 1, this is .107. Next compute

$$\hat{\zeta}_{12}/2 = \hat{p}_{21} - \hat{p}_{22} - \hat{\zeta}_{02}/3 \qquad [28]$$

which is .074. Then

$$\hat{\zeta}_{22} = \hat{p}_{11} - \hat{p}_{12} - \hat{\zeta}_{12}/2 - \hat{\zeta}_{02}/3 \quad [29]$$

which is equal to .225. Substituting these values into equation 19, the estimate of τ_{12} is $\hat{\tau}_{12} = .75$. Applying equation 6 to all seven items, it is seen that $\mu = 5.434$. In other words, it is estimated that the expected number of correct decisions is 5.434.

Next consider ρ_5. The value of S was estimated to be 16.929. From equations 20 - 26, this implies that

$$.418 \leq \rho_5 \leq .74. \quad [30]$$

This analysis suggests that these seven items, taken as a whole, are not very accurate since there is at least a 26 percent chance of making an incorrect decision on three or more items. How should the test be modified? Another important question is to what extent can it be improved? One approach to improving the test is to increase the number of distractors, and another approach is to try to modify or replace the distractors that are being used. The latter approach will be considered first.

The initial step in trying to decide whether to replace or modify the existing distractors is to determine the extent to which they can be improved. This can be done with the Δ measure in Wilcox (1981, eq. 20). This measure is just the difference between the maximum possible value of τ and the estimated value given that $\zeta_2 = \hat{\zeta}_2$. Another related measure is the entropy function (see Wilcox, 1981a). This measures the effectiveness of the distractors among the examinees who do not know the correct response by indicating the extent to which $p_2, \ldots, p_t$ are unequal. The closer they are to being equal, the more effective are the distractors, i.e., guessing is closer to being random. It has been pointed out (Wilcox, 1981a) that Δ might be objectionable as a measure of the extent to which $p_2, \ldots, p_t$ are equal, but for present purposes it would seem to be of interest because increasing ρ_k depends on the extent to which τ can be increased for each item.

Referring to Wilcox (1981a), a little algebra shows that for the case t = 3,

$$\Delta = (p_2 - p_3)/2 \ . \qquad [30]$$

For item 1 in Table 1, $\Delta = .024$, and for item 2 it is .034 (Δ is assumed to be positive, and so if $p_2 < p_3$, Δ is estimated to be zero.)

If the number of alternatives for item 1 is increased to t = 5, and if guessing is at random, then the value of τ would be .893 which represents an increase of .126 over the value of τ using the existing distractors. Thus, it would seem that one approach to improving item 1 is to find two more distractors that are about as effective as the two being used. Of course in practice, this might be very difficult to do.

Checking Certain Implications of the Model, and an Exact Test for Random Guessing

Suppose $y_1, \ldots, y_t$ have a multinomial distribution with cell probabilities $p_1, \ldots, p_t$ where $\Sigma y_i = n$ and $\Sigma p_i = 1$. This section describes an exact test of whether two or more of the p_i's are equal. In other words, the null hypothesis might be that $p_i = p_j$ for some $i \neq j$, or that $p_i = p_j = p_k$, etc. An important special case is the null hypothesis that

$$p_2 = p_3 = \cdots = p_t \qquad [31]$$

When equation 31 holds for the AUC scoring model, guessing is at random, and the distractors are performing at their maximum possible effectiveness among the examinees who do not know (see Wilcox, 1981a).

The main motivation for including this exact test in the present paper is that it is relevant when verifying certain implications of the new model described in previous sections. Consider, for example, equations 8 - 16. They imply that various inequalities must hold which includes

$$p_{11} \geq p_{12} \geq p_{13} \geq p_{23} \geq p_{33} . \qquad [32]$$

An asymptotic test of equation 32 is already available (Robertson, 1978). Suppose, however, the number of observations is moderate or small and that, for example, $p_{11} < p_{12}$ or $p_{13} < p_{23} < p_{33}$. Then to test the assumption that $p_{11} \geq p_{12}$ requires a test of $p_{11} = p_{12}$. In the second case, the null hypothesis would be $p_{13} = p_{23} = p_{33}$. Note, however, that if $p_{11} < p_{12}$ and $p_{12} > p_{13} < p_{33}$, a test of $p_{11} = p_{12}$ and $p_{23} = p_{33}$ is needed, but that $p_{12} = p_{13}$ would not be tested because $p_{12} > p_{13}$ is already consistent with equation 32.

The proposed test is based on the exact distribution of

$$S = \Sigma y_i^2 . \qquad [33]$$

An expression for the probability function of S was derived by Alam and Mitra (1981), but unfortunately their result is incorrect. (Prof. Alam has confirmed the error in a letter to the author.) A correction to the Alam and Mitra paper is in preparation which will include a correct expression for the probability function of S. To illustrate how this distribution can be used to test the implications of the model described in this paper, the distribution of S for k = 2 is given below.

Let a be the smallest integer greater than or equal to $n/2$, and let b be the largest integer between $n/2$ and n such that $b^2 + (n - b)^2 \leq s$ where s is an integer. If n is odd

$$\Pr(S \leq s) = \sum_{y=a}^{b} \binom{n}{y} p_1^y (1 - p_1)^{n-y} + \sum_{y=n-b}^{n-a} \binom{n}{y} p_1^y (1 - p_1)^{n-y} \qquad [34]$$

If n is even, subtract $\binom{n}{n/2} p_1^{n/2}(1 - p_1)^{n/2}$ from the right-hand side of equation 34.

For $k > 2$ the exact distribution of S is given by a recursive formula that will appear in a correction to the Alam and Mitra paper. To illustrate the proposed test, it is useful to also note that for $k = 3$, the joint distribution of y_2 and y_3 given y_1 is binomial with parameters $n - y_1$, $p_2/(1 - p_1)$ and $p_3/(1 - p_1)$. Thus, from equation 34,

$$\Pr(y_2^2 + y_3^2 \leq s|y_1) = \sum_{y=a}^{b} \binom{n-y_1}{y} \left(\frac{p_2}{1-p_1}\right)^y \left(\frac{p_3}{1-p_1}\right)^{n-y_1-y}$$

$$+ \sum_{y=n-y_1-b}^{n-y_1-a} \binom{n-y_1}{y} \left(\frac{p_2}{1-p_1}\right)^y \left(\frac{p_3}{1-p_1}\right)^{n-y_1-y} \qquad [35]$$

if $n - y_1$ is odd, and if $n - y_1$ is even, $\Pr(y_2^2 + y_3^2 \leq s|y_1)$ can be determined by evaluating the right-hand side of equation 35 and subtracting

$$\binom{n-y_1}{(n-y_1)/2} \left(\frac{p_2}{1-p_1}\right)^{(n-y_1)/2} \left(\frac{p_3}{1-p_1}\right)^{(n-y_1)/2} \qquad [36]$$

where $n - y_1$ replaces n in the definition of a and b.

To test the hypothesis

$$H_0: \; p_1 = p_2 = \ldots = p_k$$

compute $s = \Sigma y_i^2$ and then compute $\Pr(S \leq s)$ under the assumption that H_0 is true. If this last quantity is small, say less than α, reject H_0. Note that from Marshall and Olkin (1979, p. 391) it follows immediately that this hypothesis testing procedure is unbiased. (In other words, as the actual vector of p_i values moves "away" from H_0, the power of the test increases.)

The procedure is illustrated by testing to see whether guessing is at random on one of the items used above. The observed outcomes were $y_1 = 303$, $y_2 = 46$, and $y_3 = 21$. If guessing is at random, then, as previously indicated, $p_2 = p_3$. Since p_1 does not play a direct role in the null hypothesis, the conditional distribution of y_2 and y_3 given y_1 is used. The null hypothesis is that $p_2/(1 - p_1) = p_3/(1 - p_1) = 1/2$. Compute $s = 46^2 + 21^2 = 2116$.

$$\Pr(x_2^2 + x_3^2 \leq 2116 | y_1 = 303) \qquad [37]$$

is given by equation 35. Referring to tables compiled by Pearson (1968), equation 37 was evaluated to be .035 and so the null hypothesis would be rejected at the .05 level.

Estimating τ_{ij} When There Is Misinformation

Among the 30 items analyzed by Wilcox (in press a), the observed test scores suggest that two of the items do not conform well to the AUC scoring model described in a previous section. Thus, the proposed estimate of τ_{ij} is inappropriate. This section illustrates how this problem might be solved when a misinformation model appears to be more appropriate for some of the items on the test.

Consider a test item with t alternatives, and let ζ_t be the proportion of examinees who eliminate the correct response from consideration on their first attempt of the item. (An AUC scoring procedure is being assumed.) Once the examinee realizes that he/she has misinformation about the skill represented by the item, it is assumed that the examinee chooses the correct response on the next attempt. This assumption is made here because it seems to give a good approximation to how examinees were behaving on the items used in Wilcox (in press a). It is also assumed that if an examinee does not know and does not have misinformation, then he/she guesses at random among the t alternatives. Finally, for examinees with misinformation, assume that they believe the correct response is one of c alternatives that are in actuality incorrect. Thus, examinees with misinformation will require at least c + 1 attempts before getting the item correct. As an illustration, consider t = 5 and c = 3. Then,

$$p_1 = \zeta_{t-1} + \zeta_{t+1}/5 \qquad [38]$$

$$p_2 = \zeta_{t+1}/5 \qquad [39]$$

$$p_3 = \zeta_{t+1}/5 \qquad [40]$$

$$p_4 = \zeta_t + \zeta_{t+1}/5 \qquad [41]$$

$$p_5 = \zeta_{t+1}/5 \qquad [42]$$

where ζ_{t+1} is the proportion of examinees who do not know and who do not have misinformation.

This model gave a good fit to the observed scores in Wilcox (in press a), but an even more general model is possible. In particular, let γ be the population of examinees who have misinformation and give the correct response once they have eliminated c = 3 alternatives. Then

$$p_4 = \gamma\zeta_t + \zeta_{t+1}/5 \qquad [43]$$

$$p_5 = (1 - \gamma)\zeta_t + \zeta_{t+1}/5 \qquad [44]$$

Various modifications of the model are, of course, possible and presumably this model (with some appropriately chosen c value) will give a good fit to the observed test scores. For illustrative purposes, equations 38 - 44 are assumed to hold. The point of this section is that it is now possible to again estimate τ_{ij} where the misinformation model is assumed to hold for one or both of the items in any item pair. Note that for a single item where equations 38 - 44 hold,

$$\tau = \zeta_{t-1} + \zeta_{t+1}/t \quad . \qquad [45]$$

To estimate τ_{ij}, the joint probability of making a correct decision on a pair of items where, say, the first item is represented by a misinformation model, equation 7 must be rederived. Accordingly, let t' be the number of alternatives on the first item, and t is the number of alternatives on the second. The misinformation model assumes that on the first attempt of the item, examinees belong to one of three mutually exclusive categories, namely, they know the answer and choose it, they have misinformation and eliminate the correct response, or they do not know and guess at random. Thus, using previously established notation, equation 8 becomes,

$$p_{11} = \zeta_{42} + \zeta_{41}/2t' + \zeta_{40}/3t' + \zeta_{02}/t' + \zeta_{01}/2t' + \zeta_{00}/3t' \qquad [46]$$

where, in this illustration, t' = 5. There is no ζ_{i3} term (i = 0, 1, 2) because the misinformation model assumes that if examinees do not know, they cannot eliminate any of the distractors. More generally,

$$p_{11} = \zeta_{t'-1,t-1} + \sum_{j=0}^{t-1} \zeta_{t'-1,j}/(t - j)t' + \sum_{j=0}^{t-1} \zeta_{0j}/(t - j)t' \qquad [47]$$

Also

$$p_{k1} = p_{11} - \zeta_{42} \quad (k = 2, \ldots, t') \qquad [48]$$

$$p_{12} = \zeta_{41}/2t' + \zeta_{40} \qquad [49]$$

$$p_{1m} = \sum_{j=0}^{m} \zeta_{4j}/(t - j)t' \qquad (m = 0, \ldots, t-2) \; . \qquad [50]$$

The remaining p_{ij} values can be determined in a similar manner. For the two items being used here

$$p_{2m} = \sum_{j=0}^{m} \zeta_{0j}(t - j)t' \qquad (m = 2, \ldots, t) \qquad [51]$$

and

$$p_{3m} = p_{2m} \; .$$

The expressions for p_{4m} and p_{5m} involve the proportion of examinees who have misinformation on the first item. Let $\zeta_{t'j}$ be the proportion of examinees who have misinformation about the first item and can eliminate j distractors on the second (j = 0, ... , t-1). Previous expressions for the p_{km}'s did not involve $\zeta_{t'}$, because the misinformation model being used assumes that examinees who have misinformation will get the item correct on their fourth attempt. Of course, as previously indicated, some modification of this model (i.e., some alternative value for c) will probably be necessary when studying a different item for which there is misinformation. The point is that the p_{km}'s can be expressed in terms of the ζ_{ij}'s.

The remaining equations needed for the present situation are

$$p_{41} = \zeta_{52} + \zeta_{51}/2 + \zeta_{50}/3 + \zeta_{02}/5 + \zeta_{01}/10 + \zeta_{00}/15 \qquad [52]$$

$$p_{42} = \zeta_{51}/2 + \zeta_{50/3} + \zeta_{01}/10 + \zeta_{20}/15 \qquad [53]$$

$$p_{51} = \zeta_{02}/5 + \zeta_{01}/10 + \zeta_{00}/15 \qquad [54]$$

$$p_{52} = \zeta_{01}/10 + \zeta_{00}/15 \qquad [55]$$

$$p_{53} = \zeta_{00}/15 \; . \qquad [56]$$

Thus, starting with equation 56, ζ_{00} can be estimated by replacing p_{53} with its usual unbiased estimate, and the remaining ζ's can be estimated in a similar fashion. This, in turn, yields an estimate of τ_{ij} and so bounds on ρ_k can again be estimated as was illustrated in a previous section.

Discussion

One feature about ρ_k that might be disturbing is that generally it is an increasing function of the ζ_i's, the proportion of examinees who know the ith item. Thus, one way to ensure that ρ_k is close to one is to use easy items. This approach certainly is not being recommended. The view taken here is that the goal of the test is to determine which of n specific skills an examinee has acquired. The idea is that the student, or perhaps an entire group of students, can be given remedial work on those skills they have failed to learn. If ρ_k is small, and if it appears that adding effective distractors is difficult to do, this suggests that a conventional scoring procedure is inadequate, and that it should probably be abandoned. The possible replacements include using completion items, the AUC scoring procedure used here, or one of the many latent structure models referred to at the beginning of the paper. These models make it possible to determine whether ζ_i is small (e.g., Wilcox, in press). If it is small, perhaps all of the examinees should be given additional instruction.

The results reported in this paper might also be useful when empirically checking the assumptions of other latent structure models. For example, Macready and Dayton (1977) and Wilcox (1977) propose models where it is assumed that pairs of equivalent items are available. Two items are defined to be equivalent if examinees either know both or neither one. When equivalent items are available, the proportion of examinees who know both can be estimated (assuming local independence). Macready and Dayton checked their model with a chi-square goodness-of-fit test, but this requires at least three items that are equivalent to one another. (When there are only two items, there are no degrees of freedom left.)

For illustrative purposes, assume t=3, and consider equations 8-16. If two items are equivalent, then

$$\zeta_{21} = \zeta_{20} = \zeta_{12} = \zeta_{02} = 0 \qquad [57]$$

$$p_{12} = p_{21} = p_{22} \qquad [58]$$

$$p_{13} = p_{23} \qquad [59]$$

and

$$p_{31} = p_{23} , \qquad [60]$$

and an exact test of these equalities can be made using the procedure described in an earlier section. If one of these items is assumed to be hierarchically related to the other, again certain equalities must hold among equations 8-16, and this can again be tested (cf. White and Clark, 1973; Dayton and Macready, 1976).

Table 1

Number of Examinees Requiring i Attempts on Item 1 and j Attempts on Item 2

		Number of Attempts on Item 2		
Number of Attempts on Item 1	1	179	26	14
	2	76	8	4
	3	53	13	4

References

Alam, K., and Mitra, A. Polarization test for the multinomial distribution. Journal of the American Statistical Association, 1981, 76, 107-109.

Allen, M.J., and Yen, W.M. Introduction to measurement theory. Belmont, CA: Wadsworth, 1979.

Bergan, J.R., Cancelli, A.A., and Luiten, J.W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. Journal of Educational Statistics, 1980, 5, 65-81.

Bliss, L.B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Brownless, V.T., and Keats, J.A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Cross, L.H., and Frary, R.B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dayton, C.M., and Macready, G.B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dayton, C.M., and Macready, G.B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.

Knapp, T.R. The reliability of a dichotomous test-item: A 'correlationless' approach. Journal of Educational Measurement, 1977, 14, 237-252.

Macready, G.B., and Dayton, C.M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

References

Alam, K., and Mitra, A. Polarization test for the multinomial distribution. Journal of the American Statistical Association, 1981, 76, 107-109.

Allen, M.J., and Yen, W.M. Introduction to measurement theory. Belmont, CA: Wadsworth, 1979.

Bergan, J.R., Cancelli, A.A., and Luiten, J.W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. Journal of Educational Statistics, 1980, 5, 65-81.

Bliss, L.B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Brownless, V.T., and Keats, J.A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Cross, L.H., and Frary, R.B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dayton, C.M., and Macready, G.B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dayton, C.M., and Macready, G.B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.

Knapp, T.R. The reliability of a dichotomous test-item: A 'correlationless' approach. Journal of Educational Measurement, 1977, 14, 237-252.

Macready, G.B., and Dayton, C.M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Macready, G.B., and Dayton, C.M. The nature and use of state mastery models. Applied Psychological Measurement, 1980, 4, 493-516.

Marks, E., and Noll, G.A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

Marshall, A., and Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

Molenaar, I. On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology, 1981, 34, in press.

Pearson, K. Tables of the incomplete beta function. Cambridge: University Press, 1968.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Sathe, Y.S. Pradhan, M., and Shah, S.P. Inequalities for the probability of the occurrence of at least m out of n events. Journal of Applied Probability, 1980, 17, 1127-1132.

Weitzman, R.A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

White, R.T., & Clark, R.M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

Wilcox, R.R. New methods for studying stability. In C.W. Harris, A. Pearlman, and R. Wilcox, Achievement test items: methods of study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (a)

Wilcox, R.R. New methods for studying equivalence. In C.W. Harris, A. Pearlman, and R. Wilcox, Achievement test items: Methods of study. CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (b)

Wilcox, R.R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414. (a)

Wilcox, R.R. Recent advances in measuring achievement: A response to Molenaar. British Journal of Mathematical and Statistical Psychology, 1981, in press. (b)

Wilcox, R.R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, in press.

Wilcox, R.R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, in press. (a)

Wilcox, R.R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, in press. (b)

Wilcox, R.R. Determining the length of multiple-choice criterion-referenced tests when an answer-until-correct scoring procedure is used. Educational and Psychological Measurement, in press. (c)

DETERMINING THE LENGTH OF MULTIPLE CHOICE CRITERION-REFERENCED TESTS WHEN AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE IS USED

Rand R. Wilcox

DEPARTMENT OF PSYCHOLOGY
University of Southern California
Los Angeles, California 90007

and the

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles 90024

ABSTRACT

When determining the length of a criterion-referenced test, results in van den Brink and Koele (1980) and Wilcox (1980a, 1980b) indicate that the problem of guessing might be more serious than may have been expected. Recently, however, a new method of scoring tests was proposed that corrects for guessing without assuming guessing is at random. Moreover, empirical investigations suggest that the underlying assumptions of the new scoring procedure will frequently hold. This paper indicates how test length might be determined when the new scoring procedure is used. The results indicate that test length might be substantially reduced when the new scoring rule can be applied.

1. INTRODUCTION

Consider a single examinee and a domain of multiple choice test items. Let τ be the proportion of items the examinee knows, and let p be the examinee's percent correct true score. In criterion-referenced testing a frequent goal is determining whether an examinee's true score is above or below a known constant, say π_0. Usually the problem is formulated in terms of p (e.g., Huynh, 1976; Wilcox, 1979), but recently attention has also been given to the case where τ is the true score of interest (e.g., van den Brink and Koele, 1980; Wilcox, 1980a).

A basic problem with criterion-referenced tests is determining how many items to include on the test. Existing solutions are summarized by Wilcox (1980b). (See, also, Berk, 1980.) Although considerable progress has been made, serious problems remain. The main difficulty can be summarized briefly as follows: When the test length problem is formulated in terms of p, and a single examinee, the solution proposed by Fhanér (1974) may result in a test that is not overly long. However, if the problem is posed in terms of τ, and if guessing is assumed to be at random, van den Brink and Koele (1980) show that the test may have to be substantially longer to guarantee the same level of test accuracy as is obtained when the problem of guessing can be ignored. Wilcox (1980a) notes that the problem is much worse than indicated by van den Brink and Koele. This is not surprising because there is no particular reason to assume random guessing, and empirical studies verify that such an assumption might be unreasonable (Bliss, 1980; Cross and Frary, 1977).

Wilcox (1980b) indicates that the problem of guessing might be partially alleviated when latent structure models can be used to estimate τ, but there are clearly situations where such models are inappropriate (cf. Molenaar, 1981; Wilcox, 1981a). The result is that if multiple choice test items must be used, an unrealistically large number of items might be necessary in order to be reasonably certain of correctly classifying an examinee whose true score τ is close to the criterion score π_0.

This paper extends existing test length solutions to situations where an answer-until-correct scoring procedure can be used. An advantage of the new solution is that it corrects for guessing without assuming guessing is at random. In addition, the new results represent a substantial improvement over existing techniques when multiple-choice test items are being used.

An Answer-Until-Correct Scoring Rule

Wilcox (1981b) proposed an estimate of τ based on an answer-until-correct scoring procedure. This subsection briefly reviews the assumptions and justification for using this scoring rule.

Consider a multiple-choice test item with t alternatives, one of which is correct. An answer-until-correct test refers to situations where an examinee chooses alternatives until the correct one is identified. This is usually accomplished by having examinees erase a shield on an answer sheet until the correct alternative is chosen.

For a specific examinee and a randomly chosen item, let p_i be the probability that the correct answer is chosen on the ith attempt (i.e., the probability of i erasures is p_i). Wilcox makes certain assumptions about how an examinee behaves when attempting an item, and in terms of the p_i's, these assumptions imply that

$$p_1 \geq p_2 \geq \cdots \geq p_t \qquad [1]$$

Empirical investigations made by Wilcox (1980c, 1981b) suggest that the inequalities in equation 1 will frequently hold. For results on how to characterize n-item tests, see Wilcox, (in press). For a strong true score model, see Wilcox (1980c).

If equation 1 is assumed, a maximum likelihood estimate of τ is available via the pool-adjacent-violators algorithm (Wilcox, 1981b). Here, however, there is no loss in simply using the unrestricted maximum likelihood estimate which is

$$\hat{\tau} = (x_1-x_2)/n \qquad [2]$$

where $x_i(i=1,2)$ is the number of items for which the examinee is correct on the ith attempt (i.e., the number of times the examinee erases i shields), and n is the number of items on the test. The appeal of equation 2 is that it estimates τ without assuming guessing is at random, and as was previously noted, there is some empirical evidence that it is justified. It should be noted that $\hat{\tau}$ is theoretically justified because τ can be shown to be equal to p_1-p_2 when the assumptions in Wilcox (1981a) hold.

2. DETERMINING TEST LENGTH

This section extends the test length solutions of Fhanér (1974) and Wilcox (1979) to the answer-until-correct scoring procedure outlined above. As in Wilcox (1981b) it is assumed that x_1 and x_2 have a multinomial distribution.

Consistent with previous test length solutions (Wilcox, 1980b), the goal is to determine the smallest n so that when $\tau \leq \pi_0 - \delta^*$ or when $\tau \geq \pi_0 + \delta^*$ the probability of a correct decision (PCD) is at least P^* where $\frac{1}{2} < P^* < 1$ and $\delta^* > 0$ are predetermined constants. In this section the decision $\tau \geq \pi_0$ is made if $\hat{\tau} \geq \pi_0$; otherwise the reverse decision is reached. By convention, either decision about τ is said to be correct when τ is in the open interval $(\pi_0 - \delta^*, \pi_0 + \delta^*)$. This open interval is called the indifference zone.

When $\tau \geq \pi_0 - \delta^*$ the rule for deciding whether τ is above or below π_0 means that

$$\text{PCD} = \sum_{A} \frac{n!}{x_1!x_2!(n-x_1-x_2)} \; p_1^{x_1} \, p_2^{x_2} \, q^{x_3} \tag{3}$$

where $A = \{(x_1, x_2): \;\; x_1 - x_2 \geq n\pi_0\}$, $\Sigma x_i = n$, $q = 1 - p_1 - p_2$ and where $x_i \geq 0$ $(i=1,2,3)$. When $\tau \leq \pi_0 - \delta^*$

$$\text{PCD} = \sum_{B} \frac{n!}{x_1!x_2!(n-x_1-x_2)!} \; p_1^{x_1} \, p_2^{x_2} \, q^{x_3} \tag{4}$$

where $B = \{(x_1, x_2): \;\; x_1 - x_2 < n\pi_0\}$.

To guarantee

$$\text{PCD} \geq P^* \tag{5}$$

when $\tau \geq \pi_0 + \delta^*$ or when $\tau \leq \pi_0 - \delta^*$, we consider, as is typically done, the worst possible case. That is, the value of τ is determined that minimizes equations 3 and 4. Then the smallest integer n is found so that $\text{PCD} \geq P^*$. It follows that equation 5 is satisfied for any value of $\tau = p_1 - p_2$ not in the indifference zone.

Since the conditional distribution of x_1 given x_2 is binomial, it can be seen that for $\tau \geq \pi_0+\delta^*$, the PCD is minimized when $\tau=\pi_0+\delta^*$, and when $\tau \leq \pi_0-\delta^*$ the minimum occurs when $\tau=\pi_0-\delta^*$. Consider, for example, the case $\tau \geq \pi_0+\delta^*$. The probability of x_1 given x_2 can be written as

$$f(x_1|x_2) = \binom{n-x_2}{x_1}\left(\frac{p_2+\tau}{1-p_2}\right)^{x_1}\left(\frac{1-2p_2-\tau}{1-p_2}\right)^{n-x_1}$$

Thus, the PCD is equal to

$$\sum_{x_2=0}^{n-[n\pi_0]^+}\ \sum_{x_1=[x_2+n\pi_0]^+}^{n} f(x_1|x_2)\binom{n}{x} p_2^{x_2}(1-p_2)^{n-x_2} \qquad (6)$$

where $[x]^+$ means the smallest integer greater than or equal to x. The term $f(x_1|x_2)$ is the only one that depends on the parameter τ. Also, for each x_2, and fixed p_2, the second summation is an increasing function of τ (see, e.g., Wilcox, 1979). Thus, the value of τ that minimizes equation 6 with the restriction that $\tau \geq \pi_0+\delta^*$ is $\tau=\pi_0+\delta^*$. The case $\tau \leq \pi_0-\delta^*$ is handled in a similar fashion, and in particular, the minimum PCD occurs when $\tau=\pi_0-\delta^*$.

There remains the problem of determining the exact values of p_1 and p_2 that minimize the PCD when p_1-p_2 is equal to $\pi_0-\delta^*$ or $\pi_0+\delta^*$. An exact solution is not given, but it is possible to further limit the possible values of p_1 and then to use numerical techniques to solve the problem.

First suppose $\tau=\pi_0-\delta^*$. Since $p_1+p_2+q=1$, $2p_1+q=1+\pi_0-\delta^*$. It follows that the largest possible value for p_1 is $p_1'=(1+\pi_0-\delta^*)/2$, and because of the restriction on q, the smallest possible value is

$$p_1'' = \frac{1-\pi_0+\delta^*}{t} + \pi_0-\delta^*$$

In practice the closed interval $[p_1'', p_1']$ will be relatively short. For example, if $\pi_0=.8$ and $\delta^*=.1$, $p_1'=.85$, $p_1''=.775$. Since $p_2=p_1-\pi_0+\delta^*$ and $q=1-p_1-p_2$, the PCD can be written as a function of p_1, and the value of p_1 that minimizes the PCD can be determined.

A similar approach can be used for the case $\tau=\pi_0+\delta^*$. the lowest possible value for p_1 is

$$\frac{1-\pi_0-\delta^*}{t} + \pi_0+\delta^*$$

and the largest possible value is $(1+\pi_0+\delta^*)/2$.

Although the value of p_1 can be determined that minimizes the PCD, there will be instances where this will be inconvenient and possibly expensive to do. However, it is possible to obtain a conservative choice for n by considering the case $p_1=(1+\pi_0-\delta^*)/2$ and $q=0$. Then the PCD is equal to

$$\sum_{x=0}^{x_0-1} \binom{n}{x} p_1^x(1-p_1)^{n-x}$$

where x_0 is the smallest integer greater than or equal to $n(\pi_0+1)/2$. This situation yields a conservative value for n is the sense that for values of τ not in the indifference zone, $\hat{\tau}$ achieves its maximum variance when $p_1=(1+\pi_0-\delta^*)/2$ and $q=0$.

For this particular value of p_1, and since $q=0$, results in Fhanér (1974) and Wilcox (1979a), can be applied. In particular, an approximate solution for n is

$$n = \frac{\lambda^2(1+\pi_0)(1-\pi_0)}{(\delta^*)^2} \tag{14}$$

where λ is the P^* quantile of the standard normal distribution.

Suppose, for example, $P^*=.9$, $\delta^*=.1$ and $\pi_0=.8$. To ensure that the PCD$\geq$.9 for any τ not in the indifference zone, equation 14 says that approximately $n=59$ items are required. For $P^*=.95$, $n=97$.

Wilcox (1980b) also considered the situation where $\delta^*=.1$ and $P^*=.9$ but where the usual correction for guessing formula score was used. It was found that varying the actual probability of guessing the correct response had a substantial effect on the test length. In one instance the required test length was found to be 159, and in another it was 281. As indicated above, an answer-until-correct scoring procedure requires only 59 items without assuming guessing is at random. Thus, the results reported here are considerably more encouraging than those reported by Wilcox (1980b).

REFERENCES

Berk, R. R. A consumer's guide to criterion-referenced test reliability. Journal of Educational Measurement, 1980, 17, 323-349.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Fhanér, S. Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and and Statistical Psychology, 1980, 33, 104-108.

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement, 1979, 31, 13-22 (a).

Wilcox, R. R. An approach to measuring the achievement or proficiency of an examinee. Applied Psychological Measurement, 1980, 4, 241-251 (a).

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement, 1980, 4, 425-446 (b).

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1980, submitted for publication (c).

Wilcox, R. R. Recent advances in measuring achievement: A response to Molenaar. British Journal of Mathematical Statistical Psychology, 1981, in press (a).

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, in press (b).

A CLOSED SQUENTIAL PROCEDURE FOR COMPARING THE BINOMIAL DISTRIBUTION TO A STANDARD

Rand R. Wilcox

DEPT. of PSYCHOLOGY
UNIVERSITY OF SOUTHERN CALIFORNIA
&
CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California . Los Angeles

ABSTRACT

Fhanér (1974) proposed an approach to measuring achievement where the binomial error model is assumed, and where the goal is to determine whether an examinee's percent correct true score is above or below a known constant. Wilcox (1980b), as well as van den Brink & Koele (1980), point out that a substantially larger number of items might be required when guessing is incorporated into Fhanér's solution. The purpose of this brief note is to derive the exact sampling distribution of a closed sequential procedure that solves the problem considered by Fhanér. We then show that the probability of a correct decision under the new procedure is exactly the same as it is when Fhanér's procedure is applied. In addition, the number of observations under the closed sequential procedure is always less than or equal to the number required under the fixed sample size approach. In some cases, the number of observations is considerably less.

In the context of mental test theory, Fhanér (1974) considered the problem of comparing a binomial probability function to a standard or known constant. More specifically, it was assumed that a random variable x has a density given by

$$\binom{N}{x} \theta^{x} (1-\theta)^{N-x} \tag{1}$$

and that we want to determine whether θ is above or below a known constant θ_0. (For a recent review of the binomial error model, see Wilcox, 1981.) The goal in Fhanér's paper was to determine the minimum N so that simultaneously,

$$\sum_{x=n}^{N} \binom{N}{x} \theta^{x}(1-\theta)^{N-x} \geq P^{*}, \text{ whenever } \theta \geq \theta_o+\delta^{*} \tag{2}$$

and

$$\sum_{x=0}^{n-1} \binom{N}{x} \theta^{x} (1-\theta)^{N-x} \geq P^{*}, \text{ whenever } \theta<\theta_0-\delta^{*} \tag{3}$$

where $\frac{1}{2} < P^{*} < 1$ and $\delta^{*} > 0$ are predetermined constants, and n is an appropriately chosen passing score.

We note that in recent years, the problem considered by Fhanér has generated considerable interest in mental test theory. Wilcox (1980a) summarizes existing results.

Suppose we choose n to be the smallest integer such that $n/N \geq \theta_0$. An asymptotic solution to determining N satisfying both equations (2) and (3) is

$$N = \lambda^{2} \theta_o (1-\theta_o)/(\delta^{*})^{2} \tag{4}$$

where λ is the P^{*} quantile of the standard normal distribution (Wilcox, 1979a). From (4) it is evident that N becomes indefinitely large as δ^{*} approaches zero.

When applying Fhanér's solution to achievement tests, it may be necessary to choose δ^*small in order to take guessing into account (van den Brink & Koele, 1980; Wilcox, 1980b). This, in turn, might mean that a relatively large number of items will be required. One approach to this problem is to apply a sequential procedure, but these are optimal under circumstances that might notbe met (e.g., Wetherill, 1966). Also, depending on the values of θ and θ_o, it is possible that the number of observations will be larger when a sequential procedure is applied.

When a sequential procedure is used, it is common practice to avoid taking an inordinately large number of observations by deciding in advance the maximum number of trials that will be allowed. In this event, however, the observed number of successes is not given by the negative binomial distribution, as it ordinarily would be (e.g., Wetherill, 1966), and so we do not know the exact probability of a correct decision about whether θ has a value above or below θ_o.

For the reasons given above, we consider a closed sequential procedure for comparing θ to θ_o. First, we suppose that N and n are determined in the manner already described. Here, however, observations are assumed to be taken one at a time until there are n successes or $m=N-n+1$ failures. Let x be the number of successes and let y be the number of failures when sampling is terminated. Note that either $x=n$, in which case the possible values of y are $0,1,\ldots,m-1$; or $y=m$ and the possible values of x are $0,1,\ldots,n-1$. Our decision rule is that

$\theta \geq \theta_0$ when $x=n$; otherwise we decide that $\theta<\theta_0$. The purpose of this brief note is to show that the probability of a correct decision under this closed sequential procedure is exactly the same as it is under the fixed sample size solution proposed by Fhanér (1974). We also note that the expected number of observations for the closed sequential procedure might be substantially less than what would otherwise be required. For related results, see Alling (1966), Armitage (1957), Spicer (1962), Wald (1947), Anderson & Friedman (1960).

The Joint Distribution of x and y

Let $x_i=1$ or 0, $i=1,\ldots$ be a sequence of independent trials where $\Pr(x_i=1)=\theta$. The exact distribution of x and y can be derived as follows: If $x=n$, then by the multiplication rule of probabilities, $f(x,y|\theta)$, the joint probability of x and y, is given by

$$f(x,y|\theta) = \binom{n-1+y}{n-1} \theta^{n-1} (1-\theta)^y \cdot \theta$$

$$= \binom{n-1+y}{n-1} \theta^n (1-\theta)^y \text{ for } x=n,\ y=0,\ \ldots,\ m-1. \quad (5)$$

In a similar fashion

$$f(x,y|\theta) = \binom{m-1+x}{m-1} (1-\theta)^m \theta^x \text{ for } y=m,\ x=0,\ \ldots,\ n-1. \quad (6)$$

The relationship between the closed sequential procedure and Fhanér's fixed sample size solution.

Let

$$h(\theta) = \sum_{x=0}^{n-1} \binom{N}{x} \theta^{x} (1-\theta)^{N-x}, \quad \theta<\theta_0$$

$$\sum_{x=n}^{N} \binom{N}{x} \theta^{x}(1-\theta)^{N-x}, \quad \theta\geq\theta_0$$

In other words, for fixed N, n and any θ, $h(\theta)$ is the probability of correctly determining whether the value of θ is above or below θ_0. We show that the probability of a correct decision under the closed sequential procedure is given exactly by $h(\theta)$. That is, the accuracy of both procedures is the same, regardless of the value of θ.

Suppose the closed sequential procedure is applied and that $\theta<\theta_0$. Then the probability of a correct decision is

$$\Pr(y=m|\theta)$$

$$= \sum_{x=0}^{n-1} \binom{m-1+x}{m-1} (1-\theta)^{m} \theta^{x}$$

From Patil (1960), this is equal to

$$1- \sum_{x=0}^{m-1} \frac{(m+n-1)!}{x!(m+n-x-1)!} (1-\theta)^{x} \theta^{n+m-1-x}$$

$$= 1 - \sum_{x=0}^{m-1} \frac{N!}{x!(N-x)!} (1-\theta)^{x} \theta^{N-x}$$

$$= 1 - [1 - \sum_{x=0}^{N-m} \frac{N!}{x!(N-x)!} \theta^x(1-\theta)^{N-x}]$$

$$= \sum_{x=0}^{n-1} \frac{N!}{x!(N-x)!} \theta^x(1-\theta)^{N-x}$$

$$= h(\theta).$$

For similar reasons, $Pr(x=n|\theta) = h(\theta)$ when $\theta \geq \theta_0$. This completes the proof.

Next we note that the number of observations under the closed sequential procedure is at most N, and on the average it is less. How much less will, of course, depend on θ and θ_0. In some cases, the amount can be substantial.

Suppose, for example, N=100, $\theta_0=.8$ in which case we set n=80 and m=21. The number of observations under the sequential procedure ranges from 21 to 100. Following Fhanér (1974), suppose an indifference zone formulation of the problem is used with $\delta^*=.05$. From equation (4), an approximate lower bound to the probability of a correct decision is .894 when the fixed sample size procedure is used. The results given above indicate that the same is true when the closed sequential procedure is applied.

Figure 1 shows a plot of E(x+y), the expected number of observations using the closed sequential procedure. As is evident, for certain values of θ, E(x+y) is considerably less than 100. As already noted, because of guessing, even smaller values of δ^* might be deemed appropriate which will increase the required value for N. Thus, the closed sequential procedure might be an important and valuable tool in many situations. Figure 2 shows a plot of E(x+y) when $\theta_0=.5$, n=50 and m=51.

Concluding Remarks

The new procedure might require the same number of observations as Fhanér's, but this will be highly unlikely, particularly when N is large. On the average, the number of observations will be smaller, and in some cases, by a substantial amount. Thus, it might be possible to reduce the difficulties pointed out by Wilcox (1980b), and van den Brink & Koele (1980). Of course at least N items must be available, and any sequential procedure would seem to be inconvenient in certain situations. However, with the current interest in computerized testing, the results reported here might be useful.

We also note that for a population of examinees, our closed sequential procedure is easily extended to the empirical Bayes framework considered by Wilcox (1977, 1979b). In particiular, suppose the probability function of every examinee's observed score is given by equations (5) and (6). It is readily verified that

$$\hat{\theta} = \frac{x}{x+y}$$

is a maximum likelihood estimate of θ. Therefore, $\hat{\theta}^2$ is a maximum likelihood estimate of θ^2 (Zehna, 1966). Let $\hat{\theta}_i$ be the maximum likelihood estimate for the ith randomly sampled examinee, i=1,...,m. It follows that $M_1=m^{-1}\Sigma\hat{\theta}_i$ and $M_2=m^{-1}\Sigma\hat{\theta}_i^2$ can be used to estimate the first and second moments of the distribution of θ over the population of examinees.

If we assume the density of θ belongs to the beta family, we can also estimate test accuracy as is done by Wilcox (1977) and we can estimate test reliability in the manner described by Huynh (1976) by noting that a negative binomial density function compounded by a beta distribution yields the inverse Pólya-Eggenberger probability function (e.g., Sibuya, 1979). The details are straightforward, and so further comments are not made.

References

Anderson, T. W., & Friedman, M. (1960) A limitation of the property of the sequential probability ratio test. In I. Olkin (Ed.) Contributions to Probability and Statistics. Stanford: Stanford University Press.

Alling, D. W. (1966) Closed sequential tests for binomial probabilities. Biometrika, 53, 73-84.

Fhanér, S. (1974) Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 27, 172-175.

Huynh, H. (1976) On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Patil, G. P. (1960) On the evaluation of the negative binomial distribution with examples. Technometrics, 2, 501-505.

Sibuya, M. (1979) Generalized hypergeometric, digamma, and trigamma distributions. Annals of the Institute of Statistical Mathematics, 31, 373-390.

Spicer, C. (1962) Some new closed sequential designs for clinical trials. Biometrics, 18, 203-211.

Van den Brink, W. P., & Koele, P. (1980) Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 33, 104-108.

Wald, A. (1947) Sequential analysis. New York: John Wiley.

Wetherill, G. B. (1966) Sequential methods in statistics. London: Halsted Press.

Wilcox, R. R. (1977) Estimating the likelihood of a false-positive and false-negative decision with a mastery test: An empirical Bayes approach. Journal of Educational Statistics, 2, 289-307.

Wilcox, R. R. (1979) Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement, 39, 13-22. (a)

Wilcox, R. R. (1979) On false-positive and false-negative decisions with a mastery test. Journal of Educational Statistics, 4, 59-73. (b)

Wilcox, R. (1980) Determining the length of a criterion-referenced test. Applied Psychological Measurement, 4, 425-446. (a)

Wilcox, R. R. (1980) An approach to measuring the achivement or proficiency of an examinee. Applied Psychological Measurement, 4, 241-251. (b)

Wilcox, R. R. (1981) A review of the beta-binomial model and its extensions, Journal of Educational Statistics, 6, 3-32.

Zehna, P. W. (1966) Invariance of maximum likelihood estimation. Annals of Mathematical Statistics, 37, 744.

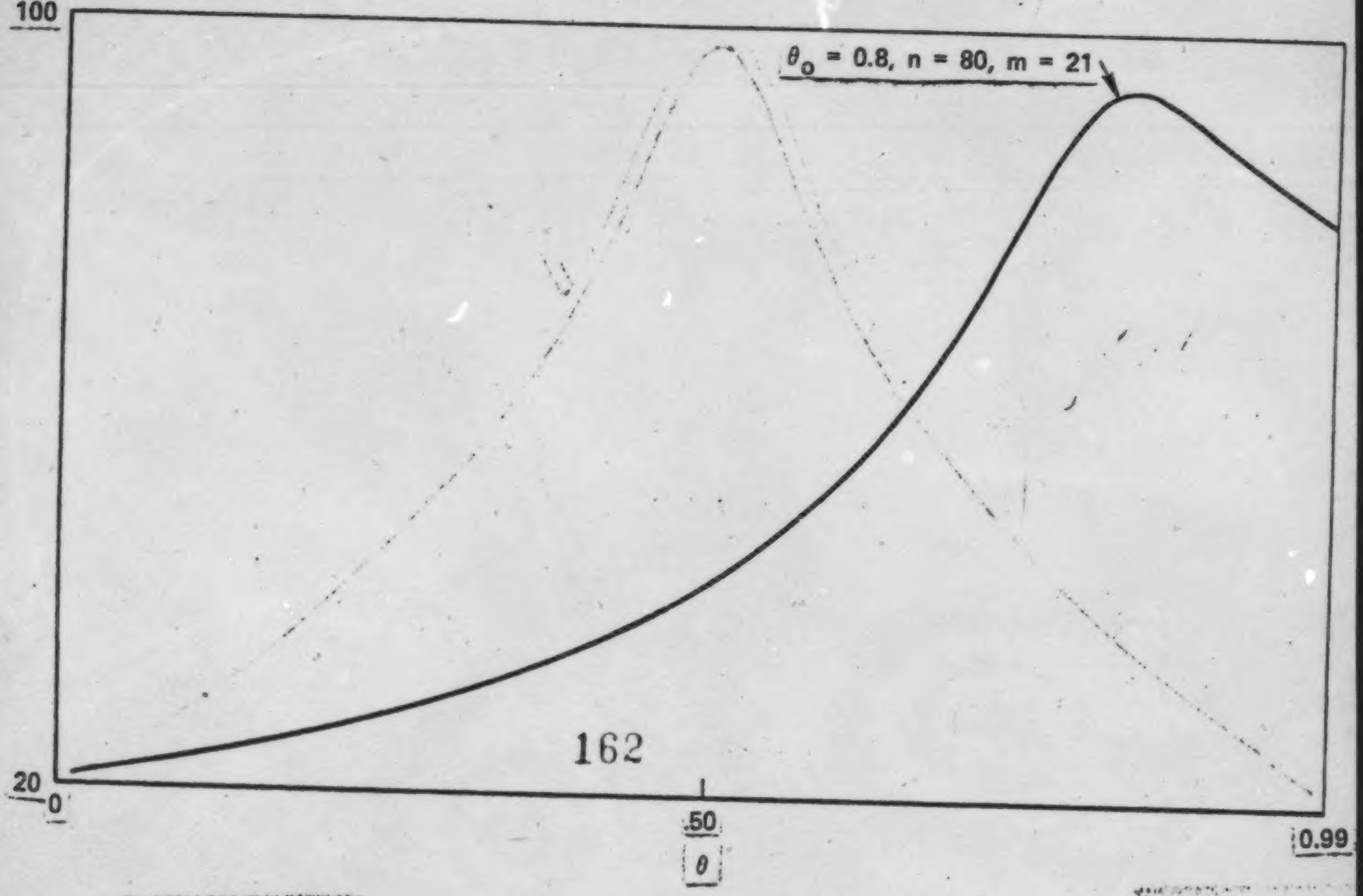

Figure 1.

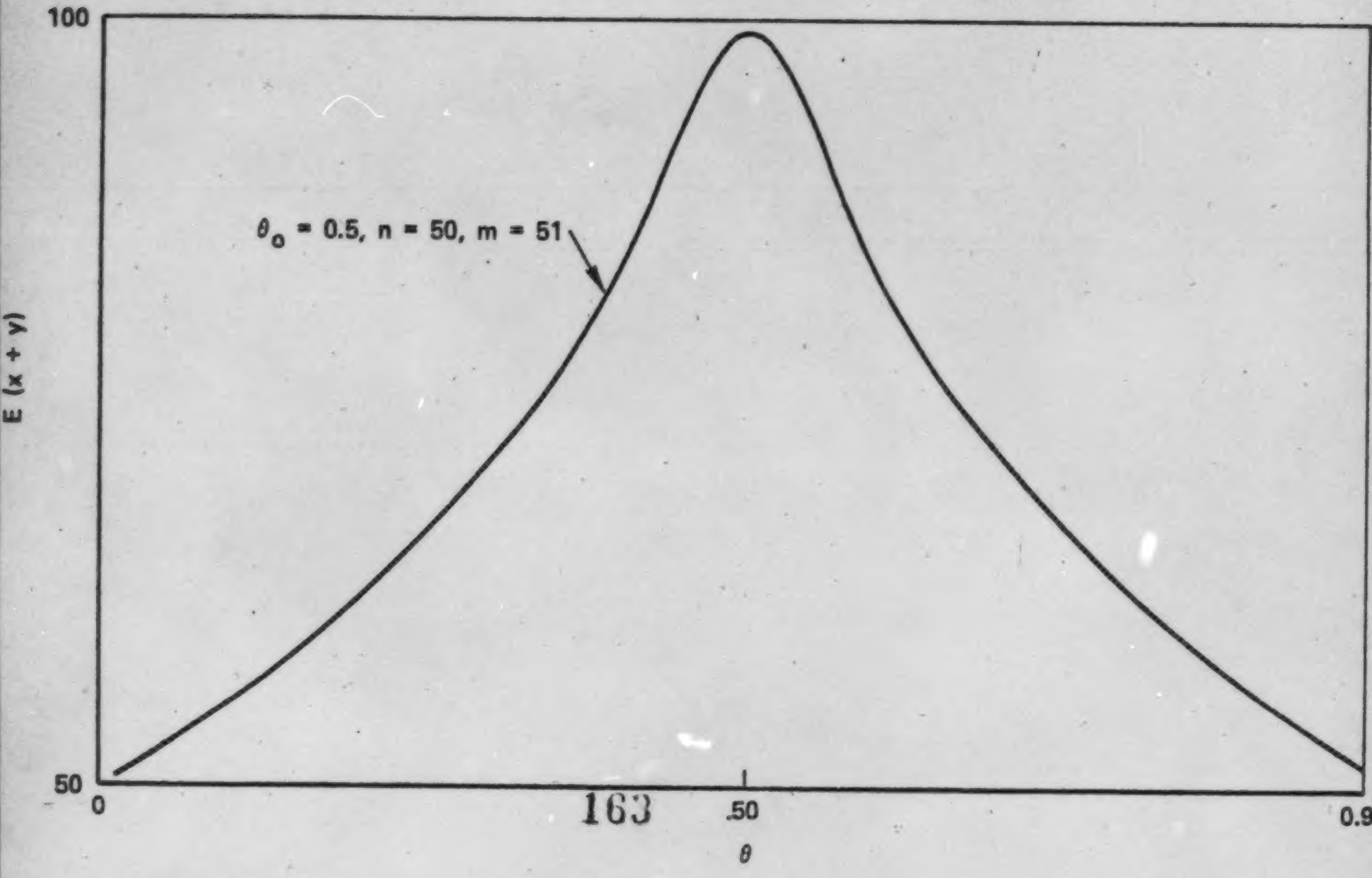

Figure 2.

A CLOSED SEQUENTIAL PROCEDURE FOR ANSWER-UNTIL-CORRECT TESTS

Rand R. Wilcox

ABSTRACT

Wilcox (1981a) proposed a latent structure model for answer-until-correct tests that can solve various measurement problems including correcting for guessing without assuming guessing is at random. This paper proposes a closed sequential procedure for estimating true score that can be used in conjunction with an answer-until-correct test. For criterion-referenced tests where the goal is to determine whether an examinee's true score is above or below a known constant, the accuracy of the new procedure is exactly the same as a more conventional sequential solution. The advantage of the new procedure is that it eliminates the possibility of using an inordinately large number of items when in fact a large number of items is not needed; typical sequential procedures always allow this possibility. In addition, the new procedure appears to compare favorably to traditional tests where the number of items to be administered is fixed in advance.

1. INTRODUCTION

Consider a multiple-choice test item with t alternatives, one of which corresponds to the correct response. Under an answer-until-correct (AUC) scoring procedure, an examinee chooses alternatives until the correct response is selected. In the past, this has been accomplished by having the examinee erase a shield on an answer sheet; the examinee knows immediately whether the correct response was chosen. If it was not, the examinee erases another shield, and this process continues until the correct alternative is chosen. Another way of administering AUC tests is with a recently developed pen that is used in conjunction with a specially treated answer sheet. The examinee marks his/her selection which causes a previously invisible mark to appear on the answer sheet. If the mark signifies an incorrect choice, another alternative is chosen. An optical scanner can then be used to count the number of attempts an examinee took on each item of the test, or, of course, the test can be scored by hand. A third way of administering AUC tests, and the one that is particularly relevant to this paper, is by computer.

AUC tests appear to have several advantages. Past investigations suggest they enhance learning (Pressey, 1950), increase reliability (Hanna, 1975; Gilman & Ferry, 1972), and under certain assumptions, they can be used to correct for guessing without assuming guessing is at random (Wilcox, 1981a). Some implications of the assumptions made by Wilcox (1981a) have been empirically investigated, and the results suggest they are frequently reasonable (Wilcox, in press, a).

The ability to measure and correct the effects of guessing is particularly important in criterion-referenced testing where the goal is to determine whether an examinee's true score is above or below a known constant (van den Brink & Koele, 1980; Wilcox, 1980). Because of guessing, an unrealistically large number of items might be required to ensure a reasonably accurate test.

The goal in this paper is to describe a closed sequential testing procedure that might be used in conjunction with Wilcox's correction for guessing formula score. The results reported here generalize those reported in Wilcox (in press, b). To help motivate the new procedure, a traditional sequential procedure is also discussed.

While the potential advantages of sequential procedures is known (e.g., Wetherill, 1975), they have the practical disadvantage of possibly requiring an even larger number of observations than would be used under a fixed sample size approach. On the average this may not happen, but there is a positive probability that a sequential procedure will need more observations. Usually this problem is avoided by deciding in advance the maximum number of observations that will be allowed under a sequential procedure, but in this case the appropriate probability function may not be known. The closed sequential procedure described below is intended to correct this problem when an AUC test is being used.

2. ASSUMPTIONS AND GOALS

This section gives a more precise description of the assumptions being made and the goals of the test.

Consider a domain of skills, and suppose every skill is represented by a multiple choice test item having t alternatives from which to choose, one of which is correct. Let τ be the proportion of skills a specific examinee has acquired, and let p_i ($i=1, \ldots, y$) be the probability that the examinee chooses the correct response on the ith attempt of a randomly chosen item. Wilcox (1981a) assumes that if the examinee has acquired the skill corresponding to a randomly sampled item, he/she gives the correct response on the first attempt. If the examinee does not know, it is assumed that at most t-2 distractors can be eliminated, and that the examinee guesses at random from among those that remain. This is, of course, an over simplification of reality since the model does not allow for misinformation, nor the possibility of knowing and inadvertantly choosing an incorrect response. Other latent structure models have been proposed that include errors at the item level such as misinformation, but these models make certain assumptions that may not hold in many situations. (See Molenaar, 1981; Wilcox, 1981a; in press b.)

Based on the above assumptions, it has been shown that $\tau=p_1-p_2$ (Wilcox, 1981a). This suggests that for an AUC scoring procedure, if there are x_1 items for which the examinee is correct on the first attempt, and if there are x_2 items for which the examinee is correct on the second attempt, τ might be estimated with

$$\hat{\tau} = (x_1-x_2)/n \qquad (2.1)$$

where n is the number of items on the test. The appeal of equation 2.1 is that it corrects for guessing without assuming guessing is at random.

Wilcox's model implies that

$$p_1 \geq p_2 \geq \cdots \geq p_t \qquad (2.2)$$

and empirical investigations suggest that this inequality will frequently be reasonable (Wilcox, in press a). Note that if (2.2) is assumed, a maximum likelihood estimate of τ, assuming x_1 and x_2 have a multinomial distribution, can be obtained via the pool-adjacent-violators algorithm (Barlow, et al., 1972) which is

$$\hat{\tau} = \begin{cases} (x_1 - x_2)/n, & x_1 \geq x_2 \\ 0, & \text{otherwise} \end{cases} \tag{2.3}$$

The two most common goals of a criterion-referenced test are estimating true score, and determining whether τ is above or below a known constant, say τ_0 (Hambleton, et al., 1978). The remainder of the paper considers these problems when a sequential or closed sequential procedure is used to estimate τ.

3. A SEQUENTIAL OR INVERSE SAMPLING PROCEDURE

This section summarizes some existing results on estimating p_i under a conventional inverse sampling procedure. The main reason for including this section is to motivate the closed sequential procedure described in section 4.

Here it is assumed that an item is randomly sampled and the examinee responds to it according to the AUC scoring procedure previously described. Once the examinee identifies the correct response, another item is randomly sampled and administered, and the process continues until there are N items for which the first alternative chosen by the examinee is the correct response. Once sampling is terminated, let y_2 be the number of items for which the examinee chooses the correct response on the second attempt of an item, and let y_3 be the number of items for which more than two attempts were needed. The probability function of y_2 and y_3 is negative multinomial which is given by

$$f(y_2,y_3|p_1,p_2) = \frac{(N-1+y_2+y_3)!}{y_2!y_3!(N-1)!} p_1^N p_2^{y_2} q^{y_3} \quad (y_2,y_3=0,1,\ldots) \quad (3.1)$$

where p_1 and p_2 are defined in section 2, and $q=1-p_1-p_2$ (e.g., Sibuya, 1964). Properties of this distribution are summarized by Sibuya (1964), Mosimann (1963), and Johnson & Kotz (1969). See, also, Olkin & Sobel (1965), Olkin (1972), and Cacoullos & Sobel (1966).

The maximum likelihood estimates of p_1 and p_2 are $\hat{p}_1=N/(N+y_2+y_3)$ and $\hat{p}_2=y_2/(N+y_2+y_3)$, respectively. As previously mentioned, $\tau=p_1-p_2$, so the maximum likelihood estimate of the examinee's true score is $\hat{\tau}=(N-y_2)/(N+y_2+y_3)$ (Zehna, 1966).

Consider the problem of determining whether τ is above or below τ_0. The obvious solution is to decide $\tau\geq\tau_0$ if and only if $\hat{\tau}\geq\tau_0$. This is the typical type of decision rule used with criterion-referenced tests, and it is the solution used here. Thus, for $\tau\geq\tau_0$, the probability of a correct decision (PCD) is

$$R = \sum_A f(y_2,y_3|p_1,p_2) \quad (3.2)$$

where $A=\{(y_2,y_3): \hat{\tau}\geq\tau_0\}$. For $\tau<\tau_0$ the PCD is just $1-R$.

Given p_1 and p_2, the PCD can be compared to the usual fixed sample size solution, and some comparisons are made in the next section. The expected number of observations is also easily computed, and it is given by $N+(p_2+q)/p_1$.

An appeal of sequential procedures is that the expected number of observations can be substantially less than what is needed under a fixed sample size approach. However, as previously indicated, there is a positive probability that the actual number of observations will be large. In practice this problem is avoided by determining in advance the maximum total number of observations that will be allowed. However, if sampling is terminated when $N+y_2+y_3$ reaches a predetermined value, the joint probability function of y_2 and y_3 is no longer given by the multinomial distribution. The next section proposes a possible solution to this problem when determining whether τ is above or below τ_0.

4. A CLOSED SEQUENTIAL PROCEDURE

Suppose the sequential procedure in section 3 is used in which case $\tau<\tau_0$ is decided if

$$\frac{N-y_2}{N+y_2+y_3} < \tau_0$$

Rearranging terms, the decision $\tau<\tau_0$ is made if

$$N(1-\tau_0) < (1+\tau_0)y_2 + \tau_0 y_3 \tag{4.1}$$

Thus, once $(1+\tau_0)y_2 \geq N(1-\tau_0)$, or $\tau_0 y_3 \geq N(1-\tau_0)$, there is no point in sampling more items because the decision $\tau<\tau_0$ will be made no matter how well the examinee performs on the remaining items.

Next suppose the inverse sampling scheme is modified so that sampling terminates when $y_1=N$, or $y_2=M$ or $y_3=m$ where y_1 is the number of items for which the examinee is correct on the first alternative~~s~~ chosen. For the moment M and m represent arbitrary integers.

The joint probability function of y_1, y_2 and y_3 can be derived in the same way as was the distribution in Wilcox (in press, b), and so the details are not given. An alternative derivation is also available by viewing the process as a random walk on a three dimensional lattice, but again the details are relatively straightforward, and so they are omitted. The result is that the joint probability function is given by

$$\frac{(N-1+y_2+y_3)!}{(N-1)!y_2!y_3!} p_1^N p_2^{y_2} q^{y_3} \quad (y_1=N, 0\le y_2<M, 0\le y_3<m) \tag{4.2}$$

$$\frac{(y_1+M-1+y_3)!}{y_1!(M-1)!y_3!} p_1^{y_1} p_2^M q^{y_3} \quad (0\le y_1<N, y_2=M, 0\le y_3<m) \tag{4.3}$$

$$\frac{(y_1+y_2+m-1)!}{y_1!y_2!(m-1)!} p_1^{y_1} p_2^{y_2} q^m \quad (0\le y_1<N, 0\le y_2<M, y_3=m) \tag{4.4}$$

The discussion of the decision rule under the sequential procedure suggests that the closed sequential solution be used with $M=N(1-\tau_0)/(1+\tau_0)$ and $m=N(1-\tau_0)/\tau_0$. If sampling terminates because $y_2=M$ or $y_3=m$ occurs, the decision $\tau<\tau_0$ is made. If sampling stops because $y_1=N$, decide $\tau\ge\tau_0$ if and only if $(N-y_2)/(N+y_2+y_3)\ge\tau_0$. This is the same decision rule used under the sequential procedure described in section 3, but this rule can be justified based solely on the probability function in equations 4.2, 4.3 and 4.4. To see this, note that the maximum likelihood estimate of $p_i (i=1,2,3)$ under the closed sequential procedure is

$$\hat{p}_i = \frac{y_i}{y_1+y_2+y_3} \tag{4.5}$$

where one and only one of the y_i's has attained its maximum value. By the choice of M and m, the decision $\tau<\tau_0$ is made if $y_2=M$ or $y_3=m$ because equation 4.5 yields an estimate of $\tau=p_1-p_2$ that is less than τ_0. If $y_1=N$, the decision $\tau\geq\tau_0$ is reached if $(N-y_2)/(N+y_2+y_3)\geq\tau_0$.

The above discussion reveals the important result that the PCD under the closed sequential procedure is exactly the same is it is under the sequential procedure. To see this, note that for $\tau\geq\tau_0$, the PCD under the closed sequential procedure is

$$\sum_A \frac{N-y_2}{N+y_2+y_3}\, p_1^{N}\, p_2^{y_2}\, q^{y_3} \qquad (4.6)$$

which is the same as expression 3.2. It follows that the PCD is also the same under the two procedures for $\tau<\tau_0$.

A Comparison of the Fixed Sample Size and Closed Sequential Solution

For a conventional item sampling model where the total number of items is fixed at n, the random variables x_1 and x_2, which were defined in section 2, have a multinomial distribution. Thus, when comparing τ to τ_0 and when $\tau\geq\tau_0$, the PCD is

$$\sum_B \frac{n!\,p_1^{x_1}\,p_2^{x_2}\,q^{n-x_1-x_2}}{x_1!\,x_2!\,(n-x_1-x_2)!} \qquad (4.7)$$

where $B=\{(x_1,x_2):\ (x_1-x_2)/n\geq\tau_0\}$. For $\tau<\tau_0$ the PCD is just one minus this quantity.

To compare the fixed and closed sequential procedure, the PCD was calculated for $n=14$, $N=10$, $\tau_0=.7$, $p_1=.85$ and $.075\leq p_2\leq.15$. This interval for p_2 was used because it is consistent with the assumption in equation 2.2 when $p_1=.85$. The results are shown in Figure 1 where the curve PS and P are the PCD under the closed sequential and fixed sample size procedure, respectively. As can be seen, the closed sequential procedure is consistently better. As an additional comparison, the PCD was computed for $p_1=.7$ and $.15 \le p_2 \le .30$. The results are plotted in Figure 2 and again the closed sequential procedure is consistently better.

CONCLUDING REMARKS

It has not been shown that the closed sequential procedure will always improve upon the fixed sample size approach to criterion-referenced tests when Wilcox's answer-until-correct scoring procedure is used. However, all indications are that given n, we can choose N, M and m so that the number of observations under the closed sequential procedure will be at most n, and yet it will give superior results. Moreover, the expected number of observations will be less. Thus, in situations where computerized testing is feasible, it would seem that the closed sequential procedure should be given serious consideration.

REFERENCES

Barlow, R., Bartholomew, D., Brem~~m~~ner, J., & Brunk, H. *Statistical inference under order restrictions*. New York: Wiley, 1972.

Cacoullos, T., & Sobel, M. An inverse sampling procedure for selecting the most probable event in a multinomial distribution. *Multivariate Analysis: Proceedings of an International Symposium*. (P.R. Krishmuiah, ed.) Academic Press, New York, 1966.

Gilman, D. A., & Ferry, P. Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*, 1972, *9*, 205-207.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-47.

Hanna, G. S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, 1975, *12*, 175-178.

Johnson, N., & Kotz, S. *Discrete distributions*. New York: Wiley, 1969.

Molenaar, I. On Wilcox's latent structure model for guessing. *British Journal of Mathematical and Statistical Psychology*, 1981, *34*, in press.

Mosimann, J. E. On the compound negative multinomial distribution and correlations among inversely sampled pollen counts. *Biometrika*, 1963, *50*, 47-54.

Olkin, I., & Sobel, M. Integral expressions for tail probabilities of the multinomial and negative multinomial distributions. *Biometrika*, 1965, *52*, 167-179.

Pressey, S. L. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. The Journal of Psychology, 1950, 29, 419-447.

Sibuya, M., Yoshimura, I., & Shimizu, R. Negative multinomial distribution. Annals of the Institute of Statistical Mathematics, 1964, 16, 409-426.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

Wetherill, G. B. Sequential methods in statistics. London: Halsted Press, 1975.

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement. 1980, 4, 425-446.

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, in press. (a)

Wilcox, R. R. Recent advances in measuring achievement: A response to Molenaar. British Journal of Mathematical and Statistical Psychology, 1981, in press. (b)

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, in press. (a)

Wilcox, R. R. A closed sequential procedure for comparing the binomial distribution to a standard. British Journal of Mathematical and Statistical Psychology, in press. (b)

Wilcox, R. R. Determining the length of multiple-choice criterion-referenced tests when an answer-until-correct scoring procedure is used. Educational and Psychological Measurement, in press. (c)

Zehna, P. W. Invariance of maximum likelihood estimation. Annals of Mathematical Statistics, 1966, 37, 744.

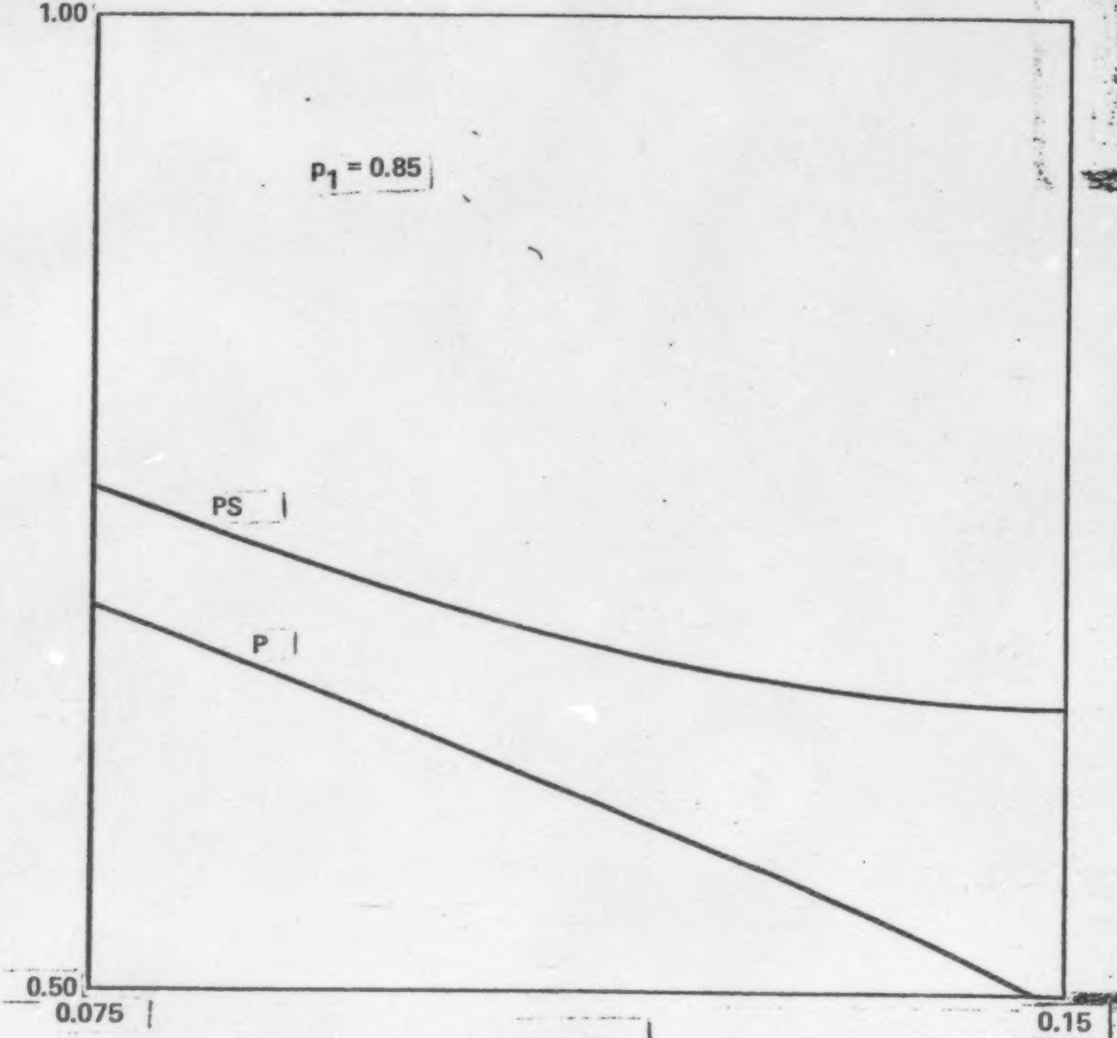

Figure 1.

.25088

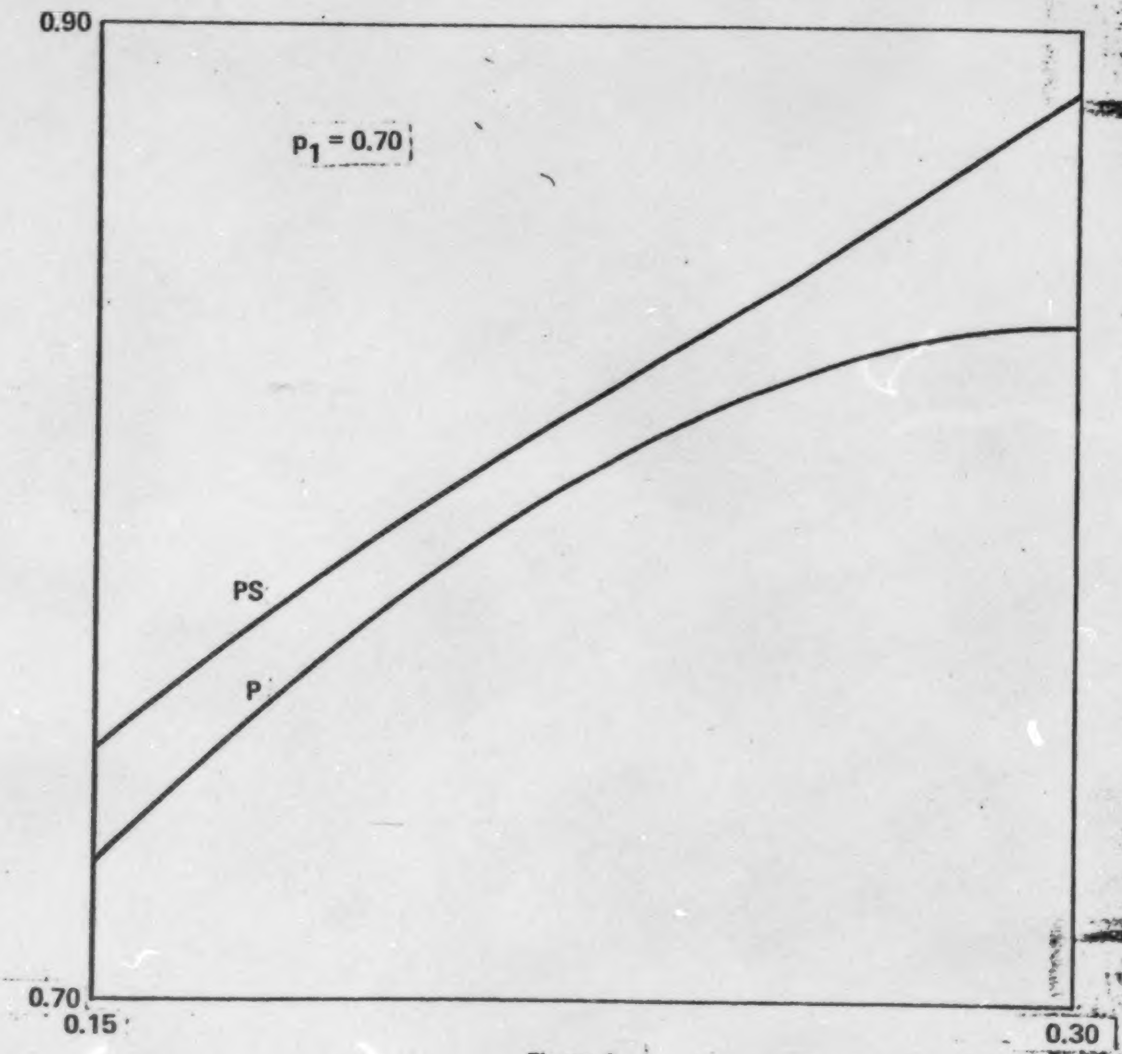

Figure 2.

Approximating the Probability of Identifying the Most Effective Treatment for the Case of Normal Distributions Having Unknown and Unequal Variances

Rand R. Wilcox

Department of Psychology
University of Southern California
and
Center for the Study of Evaluation
University of California, Los Angeles

ABSTRACT

When comparing k normal populations, an investigator might want to know the probability that the population with the largest population mean will have the largest sample mean. Put another way, what is the probability of correctly identifying the most effective treatment? The paper describes and illustrates methods of approximating this probability when the variances are unknown and possibly unequal. The results described here can also be used to measure the extent to which the populations differ for one another.

Consider k normal distributions with means μ_i and variances σ_i^2 $(i=1,\ldots,k)$. In psychology and education it is common practice to test the hypothesis that $\mu_1=\ldots=\mu_k$. If the null hypothesis is rejected, there are many instances when an investigator wants to determine which of the distributions has the largest mean. If for example, three methods of treating depression are being compared, or perhaps three methods of teaching statistics, an investigator might start by testing whether the population means are equal. If the null hypothesis is rejected, interest shifts to determining the most effective method. The obvious choice is the treatment with the largest sample mean. Once a treatment has been selected as the one most effective, it is only natural to want to determine the probability that the most effective treatment was indeed selected, i.e., we want to determine the probability that the distribution with the largest population mean will have the largest sample mean. Note that if this probability were known exactly, we would have a measure of the extent to which the treatments differ from one another (cf. Hays, 1973, pp. 481-491, Cleveland & Lachenbruch, 1974).

Typically, the approach to the problem just described is from the point of view of designing an experiment (e.g., Gibbons, Olkin, & Sobel, 1977). In particular, procedures have been devised for determining how many observations are needed so that an investigator can be reasonably certain that the most effective treatment is identified. The normal case has been considered by Bechhofer (1954), Bechhofer, Dunnett and Sobel (1954) and Dudewicz and Dalal (1975). These solutions are similar to determining power, but there are important differences.

Also, these solutions are highly conservative in the sense that if $\mu_{[k]}-\mu_{[k-1]} \geq \delta^*$ the probability of a correct selection is at least P^* where $\mu_{[k]} \geq \cdots \geq \mu_{[1]}$ are the population means written in descending order and where δ^* and P^* are predetermined constants. The value of δ^* represents the smallest difference between $\mu_{[k]}$ and $\mu_{[k-1]}$ the experimenter believes worth detecting. In actuality the difference $\mu_{[k]}-\mu_{[k-1]}$ might be considerably larger than δ^* in which case fewer observations are really needed to guarantee that the best treatment is selected for use.

Recently, Tong (1978) proposed an adaptive sequential approach to the problem of identifying the most effective treatment for the case of normal distributions having a common known variance. The motivation for the procedure is to take advantage of situations where $\mu_{[k]}-\mu_{[i]}$ $(i=1,\ldots,k-1)$ is large. The basic idea is that if the population means are substantially different fewer observations are needed than when the differences are small, say equal to δ^*. A crucial step in applying this solution is estimating the probability that the distribution with the largest population mean will produce the largest sample mean. A method of estimating this value is available, but it requires numerical quadrature which can be rather expensive to use. Accordingly, Tong uses bounds on this probability (Olkin, Sobel, & Tong, 1976) that are easily computed. The purpose of this paper is to describe, and illustrate methods of estimating similar bounds when the variances are unknown and unequal.

Description of the Procedure

Let x_{ij} ($i=1,\ldots,k$; $j=1,\ldots,n+1$) be $n+1$ randomly sampled observations from the ith normal distribution. Compute $\bar{x}_i = \sum_{j=x}^{n} x_{ij}/n$, $s_i^2 = \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2/(n - 1)$. For technical reasons explained below, it is necessary to assume that $n+1 \geq s_i^2$. This is not a serious restriction in practice since the possible values of x_{ij} are usually bounded. If, for example, there are known constants a and b such that $0 \leq a \leq x_{ij} \leq b$, and if every x_{ij} is divided by $a+b$, s_i^2 will be less than one, and the results described below can be applied.

Next compute

$$\tilde{x}_i = \sum_j a_{ij} x_{ij} \tag{1}$$

where

$$a_{i,n+1} = 1 - nc_i$$

$$a_{i1} = a_{i2} = \ldots = a_{in} = c_i$$

and

$$c_i = \frac{n + \sqrt{n^2 - n(n+1)(1 - s_i^{-2})}}{n(n+1)}$$

For technical reasons Dudewicz and Dalal select the treatment with the largest $\tilde{x}_i$ value as the one that has the largest mean. In practice this will usually be the same as selecting the treatment with the largest sample mean.

If $\tilde{x}_{(i)}$ is the value of (1) for the population having mean $\mu_{[i]}$, then the probability of a correct selection (PCS) is the probability that the distribution with mean $\mu_{[k]}$ will have the largest $\tilde{x}_i$ value. This probability is given by

$$\Pr(\tilde{x}_{(i)} \le \tilde{x}_{(k)},\ i=1,\ldots,k-1)$$

$$= \Pr(\tilde{x}_{(i)} - \mu_{[i]} \le \tilde{x}_{(k)} - \mu_{[k]} + \delta_i,\ i=1,\ldots,k-1) \qquad (2)$$

where $\delta_i = \mu_{[k]} - \mu_{[i]}$ $(i=1,\ldots,k-1)$. From Dudewicz and Dalal (1975, p.38), $\tilde{x}_i - \mu_{[i]}$ has a t distribution with $\nu=n-1$ degrees of freedom. Thus, (2) is equal to

$$\int_{-\infty}^{\infty} \prod_{i=1}^{k-1} F_\nu(z + \delta_i) f_\nu(z)\,dz \qquad (3)$$

where F_ν and f_ν are the cumulative distribution and density function, respectively, of a t distribution with ν degrees of freedom. (In Dudewicz and Dalal's notation we are setting $h=\delta^*=1$.)

From a theoretical point of view expression (3) follows from Theorem 4.1 in Dudewicz and Dalal which assumes that a two-stage sampling procedure is being used. In the first stage n observations are taken, and the second stage consists of taking $n_i - n$ additional observations sampled from the ith normal population where

$$n_i = \max\,[n + 1,\ s_i^2]$$

A slightly more general expression for c_i is also required, namely,

$$c_i = \frac{n_i - 1 + \sqrt{(n_i - 1)^2 - (n_i - 1)n_i(1 - s_i^{-2})}}{(n_i - 1)n_i}$$

In many situations a two-stage sampling procedure may be expensive or impractical, and so we have outlined how this problem might be avoided. However, when sampling is from a truly normal distribution, a two-stage procedure must be used in conjunction with the more general expression for c_i.

To estimate the probability of identifying the most effective treatment, i.e., the probability that the population with mean $\mu_{[k]}$

produced the largest $\tilde{x}_i$ value, simply replace δ_i in equation (3) with $\delta_i = \bar{x}_{[k]} - \bar{x}_{[i]}$ where $\bar{x}_{[i]}$ is the sample mean corresponding to the population that produced the ith largest $\tilde{x}_i$ value.

Bounds on the Probability of a Correct Selection

So far, nothing particularly new or unusual has been described; we have merely followed the developments in Olkin, Sobel, and Tong (1976). The only difference is that the procedure in Dudewicz and Dalal (1975) was used to handle the unknown and possibly unequal variances; Olkin et al. assume the variances are known. The main concern in this section is evaluating (3). This can be done with numerical quadrature techniques (e.g., Dudewicz, Ramberg, & Chen, 1975), but this can be expensive, particularly when the degrees of freedom are small. Accordingly, we derive upper and lower bounds on (3).

Our main result can be described as follows: Let

$$p_i = \int_{-\infty}^{\infty} F_\nu(z + \delta_i) f_\nu(z)\,dz$$

$$q_i = 1 - p_i$$

$$q_{ij} = 1 + p_i p_j - p_i - p_j$$

$$Q_1 = \sum_{i=1}^{k-1} q_i$$

$$Q_2 = \max_j \sum_{i \neq j} q_{ij}$$

, where the summation is from 1 to k~~-1~~.

Values of the integral in the definition of p_i are given in a table in Dudewicz and Dalal (1975, p. 52). Recalling that the PCS is given by (3), it will be shown that

$$PCS \geq 1 - Q_1 + Q_2 \tag{4}$$

To establish (4), the following definition is required. Let $\underline{A}=(a_1,\ldots,a_k)$ and $\underline{B}=(b_1,\ldots,b_k)$ be any two vectors, and let $a_{[1]}\geq a_{[2]}\geq\cdots\geq a_{[k]}$ and $b_{[1]}\geq b_{[2]}\geq\cdots\geq b_{[k]}$ be the components of $\underline{A}$ and $\underline{B}$ written in ascending order. A function ϕ is Schur-concave if

$$\sum_{i=1}^{r} a_{[i]} \leq \sum_{i=1}^{r} b_{[i]} \qquad \text{for } r=1,\ldots,k-1$$

and

$$\sum_{i=1}^{k} a_i = \sum_{i=1}^{k} b_i$$

implies that

$$\phi(\underline{A}) \geq \phi(\underline{B})$$

(e.g., Marshall & Olkin, 1979).

From Theorem 6.2.5 and Corollary 1 in Tong (1980, pp. 110-111) we have that $\prod_{i=1}^{k-1} F_n(z+\delta_i)$ is a Schur-concave function of the δ_i's which implies that (3) is Schur-concave as well. Thus, an upper bound to (3) is

$$\int_{-\infty}^{\infty} F_\nu^{k-1}(z+\bar{\delta})f_\nu(z)dz \tag{5}$$

where $\bar{\delta}=\sum_{i=1}^{k-1}\delta_i/(k-1)$. The integral in (5) can be evaluated via the tables in Dudewicz and Dalal (1975).

From Kimball (1951) a lower bound to (3) is

$$\prod_{i=1}^{k-1} \int_{-\infty}^{\infty} F_\nu(z+\delta_i)f_\nu(z)dz \;. \tag{6}$$

But Theorem 7.1.4 in Tong (1980, p. 147) implies that

$$\text{PCS} \geq 1 - Q_1 + \max_{j} \sum_{i\neq j} \int_{-\infty}^{\infty} F_\nu(z+\delta_j)F_\nu(z+\delta_i)f_\nu(z)dz \;.$$

Applying (6) to the summation in this last inequality establishes (4). For certain refinements of (6), see Olkin, Sobel, and Tong (1976).

Some Illustrations

To illustrate how the bounds on the PCS compare to the actual value, Monte Carlo techniques were used to evaluate (3) using arbitrarily chosen δ_i values. Column 1 in Table 1 shows the resulting approximations to (3) based on 2,000 iterations. Our computer program was checked by approximating some of the values in the tables reported by Dudewicz and Dalal (1975).

Table 1 suggests that when the value of (3) is relatively small, the upper bound given by (5) will be fairly close to the value of (3). More importantly, when (3) has a value close to one, the bounds given by (4), (5), and (6) yield a reasonably short interval which contains (3). The implication is that if, for example, we want to know whether the estimated PCS is at least .95, (4), (5), and (6) may give a fairly good indication of whether this is true.

As a final illustration, we reanalyze some data in Winer (1971, p. 153). The goal was to compare three methods of teaching, and there were 8 observations for each group. The observed scores are shown in Table 2.

Using the first seven observations in each group, we find that $c_1=.2855$, $c_2=.2959$, and $c_3=.2852$. Thus, $\tilde{x}_1=-.7060$, $\tilde{x}_2=4.112$, and $\tilde{x}_3=6.148$, and so according to the procedure in Dudewicz and Dalal, method 3 would be chosen as the most effective. (It is readily verified that $s_i^2<7$ for i=1, 2 as was required.) The question arises as to how certain we can be that method (3) is indeed the best. Since $\bar{x}_1=4.75$, $\bar{x}_2=4.625$, and $\bar{x}_3=7.75$, we have that $\delta_1=3.0$ and $\delta_2=3.125$. From a table in Dudewicz and Dalal (p. 53), the value of (5) is approximately .93. The lower bounds given by (4) and (6) are both .925. Thus, in this particular instance, we have a very good approximation to the estimated PCS. If an investigator wants the PCS to be even higher, the data indicates that additional observations must be taken.

Concluding Remarks

It is possible to sequentially estimate the PCS by applying the procedure described here in the manner proposed by Tong (1978). Many of Tong's theoretical results extend immediately to the present situation, and so further comments are omitted.

Another point is that there are alternative choices for the c_i values (e.g., Dudewicz, Ramberg, & Chen, 1975), but at present there seems to be no compelling reason for choosing one procedure over another. For a third possible procedure, see Bishop and Dudewicz (1978).

Henery (1981) proposed a method of estimating the PCS when the distributions are normal with a common known variance. We checked the accuracy of this procedure by approximating various values in the tables reported by Bechhofer (1954) -- similar checks were not made by Henery. We got reasonably good results for k=2,3 and when the PCS was less than or equal to .82, but otherwise the approximation was very poor. Despite this negative finding, a modification of Henery's procedure was tried on the case of unknown and unequal variances, but there is no indication that it would ever have any practical value. At the moment, the best approach seems to be to use the bounds on the PCS given by (4), (5), and (6).

Finally, as alluded to earlier, the results given here can be used to measure the extent to which k normal populations differ from one another. If $\mu_1=\mu_2= \ldots =\mu_k$, the PCS is equal to k^{-1} its mimimum possible value. As the δ_i values increase, so does the PCS (cf. Hedges, 1981).

References

Bechhofer, R.E. A single-sample multiple decision procedure for ranking means of normal populations with known variances. Annals of Mathematical Statistics, 1954, 25, 16-39.

Bechhofer, R.E., Dunnett, C.W., & Sobel, R. A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. Biometrika, 1954, 41, 170-176.

Bishop, T., & Dudewicz, E. Exact analysis of variance with unequal variances: Test procedures and tables. Technometrics, 1978, 20, 419-430.

Cleveland, W.J., & Lachenbruch, P.A. A measure of divergence among several populations. Communications in Statistics, 1974, 3, 201-211.

Dudewicz, E.J., Ramberg, J.S., & Chen, H.J. New tables for multiple comparisons with a control (unknown variances). Biometrische Zeitschrift, 1975, 17, 13-26.

Dudewicz, E.J., & Dalal, S.R. Allocation of observations in ranking selection with unequal variances. Sankhya, 1974, Series B, 37, 28-78.

Gibbons, J.D., Olkin, T., & Sobel, M. Selecting and ordering populations: A new statistical methodology. New York: Wiley, 1977.

Hays, W. Statistics for the social sciences. New York: Holt, Rinehart and Winston, 1973.

Hedges, L.V. Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics, 1981, 6, 107-128.

Henery, R.J. Permutation probabilities as models for horse races. Journal of the Royal Statistical Society, 1981, Series B, 43, 86-91.

Kimball, A.W. On dependent tests of significance in the analysis of variance. Annals of Mathematical Statistics, 1951, 22, 600-602.

Marshall, A.W., & Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

Olkin, I., Sobel, M., & Tong, Y.L. Estimating the true probability of correct selection for location and scale selection for location and scale parameter families (Technical Report No. 110). Stanford University, Department of Statistics, 1976.

Tong, Y.L. An adaptive solution to ranking and selection problems. The Annals of Statistics, 1978, 6, 658-672.

Tong, Y.L. Probability inequalities in multivariate distributions. New York: Academic Press, 1980.

TABLE 1

Illustrative Bounds on (3) for n = 10.

Approximate Value of (3)	δ Values					Value of (4)	Value of (5)	Value of (6)
.470	.5	1.0	1.0			.306	.470	.348
.508	.5	1.0	1.5			.364	.528	.391
.981	3.6	4.2	5.1			.977	.985	.976
.968	3.3	4.1	4.2			.964	.974	.965
.592	1.1	1.4	1.6	1.7		.407	.600	.459
.815	2.0	2.1	2.7	2.9		.747	.830	.755
.991	4.3	4.7	5.1	5.9		.988	.992	.988
.827	1.7	2.8	3.4	3.5	3.9	.790	.895	.792

TABLE 2

Method 1	Method 2	Method 3
3	4	6
5	4	7
2	3	8
4	8	6
8	7	7
4	4	9
3	2	10
9	5	9

A CAUTIONARY NOTE ON ESTIMATING THE RELIABILITY OF A MASTERY TEST WITH THE BETA-BINOMIAL MODEL

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California . Los Angeles

and the

DEPARTMENT OF PSYCHOLOGY
University of Southern California

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

ABSTRACT

Based on recently published papers, one might be tempted to routinely apply the beta-binomial model to obtain a single administration estimate of the reliability of a mastery test. Using real data, the paper illustrates two practical problems with estimating reliability in this manner. The first is that the model might give a poor fit to data which can seriously affect the reliability estimate, and the second is that inadmissible estimates of the parameters in the beta-binomial model might be obtained. Two possible solutions are described and illustrated.

1. INTRODUCTION

In recent years, efforts have been directed toward deriving ways of studying and characterizing mastery and criterion-referenced tests. A summary of the statistical and psychometric techniques that have evolved can be found in the 1980 special issue of Applied Psychological Measurement (see, also, Hambleton, et al., 1978). One approach that has received considerable attention can be described as follows: Suppose two randomly parallel test forms both consist of n dichotomously scored items. For a randomly sampled examinee, let x and y be the observed scores on the two test forms, and let f(x,y) be the joint probability function of x and y for the population of examinees. If the same passing score, say x_0, is used on both test forms, the proportion of agreement is defined to be

$$P = \sum_{x=x_0}^{n} \sum_{y=x_0}^{n} f(x,y) + \sum_{x=0}^{x_0-1} \sum_{y=0}^{x_0-1} f(x,y) . \qquad (1)$$

Many other methods have been proposed for characterizing mastery tests, but at a minimum we want P to be reasonably close to one.

Frequently it is difficult to administer two randomly parallel tests to a random sample of examinees. Accordingly, efforts have been made to derive an estimate of P based on the observed scores of only one test form. A general approach to this problem is as follows: For a specific examinee, assume the probability of an observed score x is $f(x|\theta)$, where θ is some unknown parameter, possibly vector valued. For the randomly parallel test, let $f(y|\theta)$ be the probability of an observed y, and suppose $f(x|\theta)$ and $f(y|\theta)$ are independent and they have the same parametric form. If $g(\theta)$ is the density function of θ over the population of examinees, then

$$f(x,y) = \int f(x|\theta)f(y|\theta)g(\theta)d\theta. \tag{2}$$

Once a specific form for $f(x|\theta)$ and $g(\theta)$ is assumed, it is frequently possible to estimate $g(\theta)$ which yields an estimate of $f(x,y)$. This in turn, yields an estimate of P via equation (1).

In the statistical literature, the single administration estimate of P describe above is known as an empirical Bayes approach to prediction analysis. For general results on prediction analysis, see Aitchison and Dunsmore (1975). See Geisser 1975 Sankhya.

Huynh (1976) has given a detailed account of how to estimate P for the special case where $f(x|\theta)$ (and $f(y|\theta)$) are assumed to be binomial, and where $g(\theta)$ is assumed to belong to the beta family of distributions. Note, however, that Huynh concentrates on estimating Cohen's kappa (Cohen, 1960), rather than P, once the estimate of $f(x,y)$ is available (cf. Divgi, 1980). Since Huynh's paper, several investigations of the beta-binomial model have been reported that are relevant to estimating reliability via equation (2). For example, Subkoviak (1978) compared it to three other estimates of P and concluded that all four methods gave good results, but that the beta-binomial model seemed to be the best for general use. Additional empirical support for the beta-binomial model can be found in Gross and Shulman (1980). For further results and comments on P, see Algina and Noe (1978), Huynh (1979), Divgi (1980), Traub and Rowley (1980), and Subkoviak (1980). For a recent review of the beta-binomial model, see Wilcox (1981).

Based on the studies cited above, one might be tempted to routinely apply the beta-binomial model when estimating the proportion of agreement or some related coefficient such as Cohen's kappa. In practice, though, there are at least two practical problems that might arise. First, the beta-binomial model might give a poor fit to the data (Keats, 1964a) which, as illustrated below, might affect the estimate of P. Second, the estimate of the parameters in the beta-binomial model might be inadmissible. That is, they might be negative even though the model assumes they are positive. Negative estimates can occur even when the model holds, or they might occur because the model is completely inappropriate. In some instances it might be possible to correct this problem by replacing the estimates used by Huynh (1976) with the approximation to maximum likelihood estimates described by Griffiths (1973). However, Griffiths iterative estimation procedure might not correct the problem since it can converge to inadmissible estimates even when the model holds (Wilcox, 1979). The purpose of this paper is to describe and illustrate a partial solution to these two problems.

2. TWO ALTERNATIVES TO THE BETA-BINOMIAL MODEL

Temporarily consider a single examinee responding to n dichotomously scored items. The binomial error model assumes that

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \qquad (3)$$

This assumption is theoretically justified when items are randomly sampled from an infinite item pool (or a finite pool with replacement), the examinee's responses are independent from one another, and the probability of a correct response is θ for every randomly sampled item. In many instances items are not randomly sampled, and even when they are, it is customary for every examinee to respond to the same n items. Thus, it is not surprising to find situations where (3) gives unsatisfactory results.

When trying to find a probability function that gives a good fit to data, probably three of the best known and most frequently employed distributions are the binomial, Poisson and negative-binomial (Johnson and Kotz, 1969). Thus, when the beta-binomial model is unsatisfactory, it is reasonable to consider replacing (3) with a Poisson or negative-binomial distribution. Of course, the Poisson distribution is not new to psychometric theory (Lord and Novick, 1968, chapter 21), and it frequently gives good results when a particular event occurs infrequently. The negative-binomial distribution is usually the first choice when the Poisson distribution is believed to be inadequate (Johnson and Kotz, 1969, p. 125).

The Gamma-Poisson Model

Let w=n-x and z=n-y be the number of incorrect responses given by an examinee on the first and second test forms, respectively. We begin by replacing (3) with the assumption that the probability function of w, as well as z, is Poisson with parameter η. Symbolically

$$f(w|\eta) = e^{-\eta}\eta^{w}/w! \tag{4}$$

The reason for working with w and z, rather than x and y is that the data in our example is skewed to the right. If the observed frequencies had been skewed to the left, we would have used x and y.

We also assume that for the population of examinees, η has a gamma distribution. The motivation for this assumption is that it is typically made for the Poisson case, it is mathematically convenient, and it has given good results with mental test data (Wilcox, 1981). If $f(w|\eta)$ and $f(z|\eta)$ are assumed to be independent, results in Aitchison and Dunsmore (1975) tell us immediately that

$$f(w) = \frac{\Gamma(\alpha+w)}{\Gamma(\alpha)\Gamma(w+1)} \left(\frac{\beta}{\beta+1}\right)^{w} \left(\frac{1}{\beta+1}\right)^{\alpha} \tag{5}$$

i.e., the marginal probability function of w is negative binomial. The parameters α and β can be estimated as follows: Let $\bar{w}$ and s^2 be the sample mean and variance of w for a random sample of examinees. Then $\hat{\beta}=(s^2/\bar{w})-1$ and $\hat{\alpha}=\bar{w}/\hat{\beta}$ estimate β and α respectively. Three other estimates of α and β are also available (Johnson and Kotz, 1969).

Again referring to Aitchison and Dunsmore (1975), we have that

$$f(z|w) = \frac{\Gamma(\alpha+w+z)}{\Gamma(\alpha+w)\ (z+1)} \left(\frac{\beta}{2\beta+1}\right)^{z} \left(\frac{\beta+1}{2\beta+1}\right)^{\alpha+w} \tag{6}$$

Since $f(w,z)=f(w)f(z|w)$, we have an estimate of P once α and β are determined.

The Gamma Product-Ratio Poisson Model

The other model we consider also assumes (4), but η is assumed to have a "gamma product-ratio" distribution (Sibuya, 1979). In this case

$$f(w) = \frac{\Gamma(w+\alpha)\Gamma(\beta+\gamma)\Gamma(w+\beta)\Gamma(\alpha+\gamma)}{\Gamma(w+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)\Gamma(\alpha+\beta+\gamma+w)} \tag{7}$$

where α, β, $\gamma>0$ are unknown parameters. We note that two alternative names for (7) are generalized Waring and negative-binomial beta. Also, the parameters α and β in (7) are different from those in (6).

To estimate α, β and γ, we first note that the first three factorial moments are

$$\mu_1 = \alpha\beta/(\gamma-1) \tag{8}$$

$$\mu_2 = \alpha(\alpha+1)\beta(\beta+1)/[(\gamma-1)(\gamma-2)] \tag{9}$$

$$\mu_3 = \alpha(\alpha+1)(\alpha+2)\beta(\beta+1)(\beta+2)/[(\gamma-1)(\gamma-2)(\gamma-3)] \tag{10}$$

It follows that

$$\left(\frac{\mu_2}{\mu_1} - \mu_1\right) \gamma - \alpha - \beta = \frac{2\mu_2}{\mu_1} - \mu_1 + 1 \quad . \tag{11}$$

and

$$\left(\frac{\mu_3}{\mu_2} - \mu_1\right) \gamma - 2\alpha - 2\beta = \frac{3\mu_3}{\mu_2} - \mu_1 + 4 \tag{12}$$

Thus, if $\hat{\mu}_1$ is the usual estimate of μ_i $(i=1,2,3)$, we have an estimate of γ, say $\hat{\gamma}$. Substituting $\hat{\gamma}$ and $\hat{\mu}_1$ and $\hat{\mu}_2$ into equations (8) and (9) yields

$$\alpha = \hat{\mu}_1 (\hat{\gamma}-1)\beta^{-1} \tag{13}$$

$$\alpha+\beta = \frac{\hat{\mu}_2}{\hat{\mu}_1} (\hat{\gamma}-2) - \hat{\mu}_1(\hat{\gamma}-1) - 1 \tag{14}$$

Substituting the right-hand side of (13) for α in (14) yields a quadratic equation for β. In terms of the marginal density (7), either estimate of β can be used since the other estimate of β will correspond to α, and since (7) is symmetric in α and β.

Finally, to estimate P with equation (1), we note that

$$f(w,z) = \frac{\Gamma(\alpha+w)\Gamma(\alpha+z)\Gamma(\beta+\gamma)\Gamma(\beta+w+z)\Gamma(2\alpha+\gamma)}{\Gamma(\alpha)\Gamma(\alpha)\Gamma(w+1)\Gamma(z+1)\Gamma(\beta)\Gamma(\gamma)\Gamma(2\alpha+\beta+\gamma+z+w)} \tag{15}$$

One way to establish this result is to assume $f(w|\theta)$ is negative-binomial and that $g(\theta)$ is beta (which is equivalent to assuming (7)) and then perform the integration in (2).

3. NUMERICAL ILLUSTRATIONS

This section uses real data to illustrate the practical advantages of estimating P with the two alternative estimates described above.

First we consider the data reported in Keats (1964). As previously indicated, the beta-binomial model gives a poor fit to the observed test scores, but, as noted in Wilcox (1981), the gamma-Poisson model gives a reasonably good fit. The test had $n=30$ items, and Keats reports observed test scores for 1000 examinees. If we estimate P with the beta-binomial model, the results is .90. If we use the gamma-Poisson model, the estimate is .81. The third estimate of P does not apply since the estimate of the parameters in (15) are inadmissible. Note that the reliability estimates used by Subkoviak (1976) as well as Marshall and Haertel (1975) also assume the binomial error model holds. Since the beta-binomial model gives a poor fit to data, there is some doubt about whether these estimates should even be considered.

As another illustration, suppose we have an $n=15$ item test with a passing score of $x_0=10$. Further suppose we have test scores as reported in Table 1. These results are based on real data reported in Irwin (1968) but they do not represent tests scores. The point is that we might get observed frequencies that are skewed, as are the frequencies in Table 1, in which case it might be better, or even necessary to replace the beta-binomial model with something else.

For the data in Table 1, the estimates of the parameters in the beta-binomial model are negative, and so an estimate of P cannot be made. Suppose instead (7) holds. It follows that $\hat{\alpha}=5.2162$, $\hat{\beta}=1.297$ and $\hat{\gamma}=7.7967$. Thus, the estimate of P is .97. If instead we use the gamma-Poisson model, the estimate of P is again .97.

CONCLUDING REMARKS

The main point in the paper is that the beta-binomial model might give a substantially different estimate of reliability relative to some other model that gives a better fit to data. We illustrated two possible solutions, but virtually any form for $f(x|\theta)$ can be used to estimate P via equation (2) as long as an estimate of $g(\theta)$ can be obtained.

REFERENCES

Aitchison, J., & Dunsmore, I. R. Statistical prediction analysis. London: Cambridge University Press, 1975.

Algina, J., & Noe, M. J. A study of the accuracy of Subkoviak's single administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 1978, 15, 101-110.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Divgi, D. R. Group dependence of some reliability indices for mastery tests. Applied Psychological Measurement, 1980, 4, 213-218.

Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics, 1973, 29, 637-648.

Gross, A. L., & Shulman, V. The applicability of the beta-binomial for criterion-referenced testing. Journal of Educational Measurement, 1980, 17, 195-202.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics, 1979, 4, 231-246.

Irwin, J. O. The generalized Waring distribution applied to accident data. Journal of the Royal Statistical Society, 1968, 131, Series A, 205-225.

Johnson, N., & Kotz, S. Discrete distributions. New York: Wiley, 1969.

Keats, J. A. Some generalizations of a theoretical distribution of mental test scores. Psychometrika, 1964, 29, 215-231.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Marshall, J. L., & Haertel, E. H. A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. Paper presented at the Annual meeting of the American Educational Research Association, 1975.

Sibuya, M. Generalized hypergeometric, digamma, and trigamma distributions. Annals of the Institute of Statistical Mathematics, 1979, 31, 373-390.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M. Decision-consistency approaches. In R. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore: The Johns Hopkins University Press, 1980.

Traub, R., & Rowley, G. L. Reliability of test scores and decisions. Applied Psychological Measurement, 1980, in press.

Wilcox, R. R. Estimating the parameters of the beta-binomial distribution. Educational and Psychological Measurement, 1979, 31, 527-535.

Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, to appear.

ANALYZING THE DISTRACTORS OF MULTIPLE-CHOICE TEST ITEMS OR PARTITIONING MULTINOMIAL CELL PROBABILITIES WITH RESPECT TO A STANDARD

Rand R. Wilcox

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles 90024
and the
DEPARTMENT OF PSYCHOLOGY
University of Southern California
Los Angeles, California 90007

The work upon which this publication is based was performed pursuant to a grant [contract] with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

ABSTRACT

When analyzing the distractors of multiple-choice test items, it is sometimes desired to determine which of the distractors has a small probability of being chosen by a typical examinee. At present, this problem is handled in an informal manner. In particular, using an arbitrary number of examinees, the probabilities associated with the distractors are estimated and then sorted according to whether the estimated values are above or below a known constant p_0. In this paper a more formal framework for solving this problem is described. The first portion of the paper considers the problem from the point of view of designing an experiment. The solution is based on a procedure similar to an indifference zone formulation of a ranking and selection problem. A later section considers methods that might be employed in a retrospective study. Brief consideration is also given to how an analysis might proceed when a test item has been altered in some way.

KEY WORDS: indifference zone; empirical Bayes;

Consider a multiple-choice test item having k+1 alternatives from which to choose. One of these alternatives is designated as being correct and the remaining k alternatives are referred to as distractors. Henrysson (1971, pp. 136-137) suggests that a statistical analysis of the distractors might be made as follows: Administer the item to a random sample of n examinees; if the observed frequency corresponding to a particular distractor is small, perhaps it should be replaced or rewritten.

Henrysson's procedure certainly seems like a reasonable one and in fact it is often used. A proposed distractor might appear to be satisfactory but in reality it might be infrequently chosen by examinees who do not know the correct response. It is only natural then to conduct an empirical investigation to determine when this occurs. Insofar as we want to discover whether an examinee knows the correct response, rewriting or replacing the distractor might be in order when the data suggests that it is seldom chosen. The idea is to modify the distractor in the hope of lowering the probability of guessing the correct response. It should be stressed, however, that if any or all distractors are infrequently chosen, this does not necessarily mean that the distractors should be replaced. If, for example, all of the distractors are seldom chosen, it may be that most examinees know the answer in which case the item might be acceptable for certain types of achievement tests while for other situations (e.g., Lord and Novick, 1968, p.320) the item might be discarded altogether. The statistical techniques described here are merely meant to alert a test constructor to the possibility of improving the distractors.

Let p_i $(i=1,\dots,k)$ be the probability that a randomly selected examinee chooses the ith distractor. For convenience, the (k+1)-th alternative is assumed to be the correct option. Thus, p_{k+1} is the probability of a correct response by a randomly chosen examinee. Consistent with Henrysson (1971), suppose that for each distractor we want to determine whether p_i is less than or greater than some known constant p_0. If $p_i<p_0$, the value of p_i is said to be

small and consideration is given to rewriting or replacing the distractor. If $p_i \geq p_0$, no action is taken. A common value for p_0 appears to be .1 although other values are certainly possible.

Let x_i be the number of examinees who choose the ith distractor. Since x_i/n estimates p_i, a natural decision rule (and the one that is used) is to decide the $p_i < p_0$ if $x_i/n < p_0$; if $x_i/n \geq p_0$ the reverse decision is made. A correct decision for all k distractors is made if simultaneously $x_i/n < p_0$ when $p_i < p_0$ and $x_i/n \geq p_0$ when $p_i \geq p_0$ $(i=1,\ldots,k)$. The difficulty is that because of sampling fluctuations, we might observe an x_i that results in an incorrect decision. For example, we might observe $x_i/n \geq p_0$ when in reality $p_i < p_0$. Accordingly, when using Henrysson's procedure, we need to consider the following types of questions. How many examinees should we sample to be reasonably certain of making a correct decision for all k distractors regardless of the actual values of the p_i's? This type of question occurs when designing a study of a proposed item, i.e., prior to collecting any data. In contrast, once data is available, one might conduct a retrospective study and consider the probability of making a correct sort of the distractors for the "typical" item under consideration. Still another type of problem that might be considered is determining the effect of rewriting or replacing a distractor. In the present context we would want the new value of p_i, say p_i', to be greater than p_0. At a minimum, we want p_i' to be at least as large as p_i. Thus, the question might arise as to how certain we can be that p_i' is less than or greater than p_i based on the number of examinees that are sampled. If $p_i' < p_i$, the original version of the distractor should be used; if $p_i' \geq p_i$, the new version is described as improving upon the old. The purpose of this paper is to provide an approach to these problems.

From a statistical point of view this paper is concerned with comparing multinomial cell probabilities to a standard and with comparing binomial distributions to a control. For related results on this type of problem the reader is referred to Gibbons, Olkin and Sobel (1977, Chapter 10), Fhaner (1974), Huang (1975), Tong (1969) and Wilcox (1979a, 1979b).

2. Mathematical Statement of the Problem

For a random sample of n examinees (sampled from an infinite population or a finite population with replacement) let $\underline{x} = (x_1,\ldots,x_{k+1})$ be the observed frequencies among the k+1 alternatives. The random vector $\underline{x}$ has a multinomial distribution given by

$$f(\underline{x}) = n! \prod_{i=1}^{k+1} p_i^{x_i}/x_i!$$

where $\sum_i x_i = n$ and $\sum_i p_i = 1$. Let p_0 be a known constant. The first goal is to determine for each $p_i (i=1,\ldots,k)$ whether p_i is above or below p_0. As previously indicated, the decision $p_i \geq p_0$ is made if $x_i/n \geq p_0$; otherwise the reverse is said to be true. Let g, $0 \leq g \leq k$, be the number of p_i's such that $p_i \geq p_0$ and for convenience (and without loss of generality), suppose that the p_i's $(i=1,\ldots,k)$ are ordered, i.e., $p_1 \leq p_2 \leq \ldots \leq p_k$. As already noted, in terms of the x_i's, a correct decision (CD) is made if simultaneously

(2.1) $x_i/n < p_0$, $i=1,\ldots,k-g$

and

(2.2) $x_i/n \geq p_0$, $i=k-g+1,\ldots,k$.

The problem is to find the smallest n, say n_0, so that regardless of the actual values of the p_i's, the probability of a correct decision has a value reasonably close to one. More briefly, we want to find the smallest n so that

(2.3) $P(CD) \geq P^*$,

where $2^{-k} \leq P^* \leq 1$.

Following Gibbons et al. (1977), an indifference zone formulation of the problem is used. Thus, the investigator is assumed to have chosen a constant δ^* with the idea that if $p_0-\delta^* < p_i < p_0+\delta^*$, there is negligible loss in misclassifying the ith distractor. In fact, if the value of p_i is in the open interval $(p_0-\delta^*, p_0+\delta^*)$, any decision for that distractor is designated as being correct and so a correct decision is made with probability one. Thus, our only concern is with values of $p_i \leq p_0-\delta^*$ and $p_i \geq p_0+\delta^*$.

3. An Exact Solution

In this section an exact solution to the problem of determining n_0 is described. First we observe that the P(CD) is a function of the unknown p_i's. Thus, for a given n, it might be that the $P(CD) \geq P^*$ for some values of the p_i's but not for others. To be certain that (2.3) holds for any vector $\underline{p}=(p_1,\ldots,p_{k+1})$ we consider, as is typically done, the worst possible case, namely, the p_i values, say $\underline{p}^0=(p_1^0,\ldots,p_{k+1}^0)$, that minimizes the P(CD). It is shown below that $\underline{p}^0$ does not depend on n. Hence, by choosing the smallest n so that $P(CD|\underline{p}=\underline{p}^0) \geq P^*$, (2.3) is guaranteed regardless of the actual values of the p_i's. To avoid certain technical difficulties, it is assumed that $k(p_0-\delta^*) \leq 1$. This is not a serious restriction for the problem at hand since typically $p_0 \leq .2$, $.01 \leq \delta^* \leq .1$ and $k \leq 4$.

Our immediate goal is to show that $\underline{p}^0$ is given by $p_i^0=p_0-\delta^*$ $(i=1,\ldots,k-g)$ and $p_i^0=p_0+\delta^*$ $(i=k-g+1,\ldots,k)$. First, however, some preliminary results are needed. Accordingly, we begin by demonstrating that for fixed g and n,

$$(3.1) \qquad P(x_i/n < p_0,\ i=1,\ldots,k-g)$$

is minimized when $p_1=p_2=\ldots=p_{k-g}=p_0-\delta^*$. Since by assumption $(k-g)(p_0-\delta^*) \leq 1$, the possibility of having $p_1=\ldots=p_{k-g}=p_0-\delta^*$ is ensured.

Let s be the smallest integer greater than or equal to np_0 and let $\bar{p}_i = \sum_{j=i+1}^{k-g} p_j$. Olkin and Sobel (1965) show that (3.1) is equal to

$$\frac{\Gamma(n+1)}{\Gamma^{k-g}(s)\Gamma(n-s_0+1)} \int_{p_1}^{1-\bar{p}_1} \int_{p_2}^{1-t_1-\bar{p}_2} \cdots \int_{p_{k-g}}^{1-t_1-\cdots-t_{k-g-1}} (1-t_0)^{n-s_0} \prod_{i=1}^{k-g} t_i^{s-1} \prod_{i=1}^{k-g} dt_i$$

where Γ is the usual gamma function, $s_0=(k-g)s$, $t_0=\sum_{i=1}^{k-g} t_i$ and $t_0, t_1, \ldots, t_{k-g}$ are dummy variables. Note that this quantity depends only on $(p_1,\ldots,p_{k-g})$. Examination of the limits of this (k-g)-fold integral reveals that among all vectors $(p_1,\ldots,p_{k-g})$ for which $p_i \le p_0-\delta^*$, (3.1) attains its minimum value when

(3.2) $\quad p_1=\ldots=p_{k-g}=p_0-\delta^*$

as was to be shown.

Next consider

(3.3) $\quad P(x_{k-g+1} \ge np_0,\ldots,x_k \ge np_0)$.

From Olkin and Sobel (1965) we see that this probability is equal to

$$(3.4) \qquad \frac{\Gamma(n+1)}{\Gamma^{g}(s)\Gamma(n-gs+1)} \int_0^{p_{k-g+1}} \cdots \int_0^{p_k} (1-t_0)^{n-sg} \prod_{i=1}^{g} t_i^{s-1} \prod_{i=1}^{g} dt_i$$

where now $t_0 = \sum_{i=1}^{g} t_i$ and again $t_0, t_1,\ldots,t_g$ are dummy variables. From (3.4) it follows that for fixed g and n, among all possible values of $p_i \ge p_0 + \delta^*(i=k-g+1,\ldots,k)$, expression (3.3) is minimized when

(3.5) $\quad p_{k-g+1}=\ldots=p_k=p_0+\delta^*$.

The above results are now extended to show that for any n and any admissible g,

$$P(CD) = P(x<s,\dots,x_{k-g}<s,x_{k-g+1}\geq s,\dots,x_k\geq s)$$

is minimized when

(3.6) $$p_1=\dots=p_{k-g}=p_0-\delta^* \text{ and } p_{k-g+1}=\dots=p_k=p_0+\delta^*.$$

The vector $\underline{p}$ that satisfies the two conditions given by (3.6) is referred to as the least favorable configuration of the p_i's.

First note that

(3.7) $$P(CD) = \Sigma\Sigma P(x_1,\dots,x_{k-g})\ P(x_{k-g+1},\dots,x_k|x_1,\dots,x_{k-g})$$

where the first summation is over all vectors $(x_1,\dots,x_{k-g})$ such that $x_i < s$ $(i=1,\dots,k-g)$ and the second is over all vectors $(x_{k-g+1},\dots,x_k)$ such that $x_j \geq s$ $(j=k-g+1,\dots,k)$. It can be verified using standard techniques that

$$P(x_{k-g+1},\dots,x_k|x_1,\dots,x_{k-g})$$

is a multinomial distribution given by

$$\frac{(n-x_1-\dots-x_{k-g})!\ p_{k-g+1}^{x_{k-g+1}}\dots p_k^{x_k}(1-p_1-\dots-p_k)^{n-x_1-\dots-x_k}}{x_{k-g+1}!\dots x_k!(n-x_1-\dots-x_k)!\ (1-p_1-\dots-p_{k-g})^{n-x_1-\dots-x_{k-g}}}$$

$$= \frac{(n-x_1-\dots-x_{k-g})!\ r_{k-g+1}^{x_{k-g+1}}\dots r_k^{x_k}(1-p_1-\dots-p_k)^{n-x_1-\dots-x_k}}{x_{k-g-1}!\dots x_k!(n-x_1-\dots-x_k)!\ (1-p_1-\dots-p_{k-g})^{n-x_1-\dots-x_k}}$$

where $r_i=p_i/(1-p_1-\dots-p_{k-g})$.

Thus, making the appropriate modification in (3.4) or referring to Olkin and Sobel (1965) the second summation in (3.7) can be written as

$$\frac{\Gamma(n-x_1-\dots-x_{k-g}+1)}{\Gamma^g(s)\Gamma(n-x_1-\dots-x_{k-g}-gs+1)}\int_0^{r_1}\dots\int_0^{r_k}(1-t_0)^{n-x_1-\dots-x_{k-g}-sg}\prod_{j=k-g+1}^{k} t_j^{s-1}$$

$$\prod_{j=k+g+1}^{k} dt_j\ .$$

where $t_0 = \sum_{j=k+g+1}^{k} t_j$ and again $t_0, t_{k+g+1}, \ldots, t_k$ are dummy variables. Examination of the limits of this g-fold integral reveals that for fixed $x_1, \ldots, x_{k-g}$, $p_1, \ldots, p_{k-g}$, the second summation in (3.7) is minimized when $p_{k-g+1} = \ldots = p_k = p_0 + \delta^*$. This in turn implies that for fixed $p_1, \ldots, p_{k-g}$ the P(CD) as given by (3.7) is minimized when (3.5) holds.

Next, set $p_{k-g+1} = \ldots = p_k = p_0 + \delta^*$. Since by assumption it is possible to have $p_1 = \ldots = p_{k-g} = p_0 - \delta^*$, it follows, using an argument similar to the one in the preceding paragraph, that the P(CD) is minimized when (3.6) holds. Hence, by choosing n_0 to be the smallest integer such that the P(CD) $\geq P^*$ for all admissible values of g under the least favorable configuration (3.6), we guarantee (2.3) no matter what the values of the p_i's happen to be.

Exact and Approximate Methods for Calculating n_0.

Tables 1-3 give the value of n_0 for $p_0 = .1$, $\delta^* = .05$; $p_0 = .15, .2$, $\delta^* = .05, .1$; $P^* = .77, .9, .95, .99$; and k=1(1).3. If, for example, k=2, $p_0 = .1$, $\delta^* = .05$ and $P^* = .9$, n=110 examinees guarantees that the correct sort of the two distractors will be made with probability at least .9 regardless of the actual p_i values. This section describes exact and approximate methods for determining n_0.

A Lower Bound to n_0

There might be occasions where it is helpful to have a lower bound to n_0 that is easily computed. Accordingly, let I(k,p,s,n) represent the value of (3.4) when $p_1, \ldots, p_k$ have a common value p. This is also the P(CD) for the least favorable configuration when g=k. It can be seen that the smallest n, say n_1, such that $I(k,p,s,n_1) \geq P^*$ is a lower bound to n_0 whenever $p \geq p_0 + \delta^*$. Sobel, Uppuluri and Frankowski (1977) have tabled the values of $I(k,p,v,n)$ for $p=t^{-1}$, $t=k+1(1)10$, and $v=1(1)10$ which can be used to determine a lower bound to n_0 by referring to the entries for the smallest $p \geq p_0 + \delta^*$ and the largest $v < s$. For example, if $k=2$, $p_0=.1$, $\delta^*=.05$, $P^*=.9$, then the smallest $p \geq p_0 + \delta^* = .15$ in their Table B is $p = 1/6$. Examination of the entries in their table reveals that with $n=68$ (which implies that $s=7$) $I(2, 1/6, 7, 68) = .9008$ and so $n_0 \geq 68$. Thus, for this particular case, n_0 can be determined exactly by starting with $n=68$, evaluating the P(CD) for $g=0, 1, 2$ and checking whether $P(CD) \geq P^*$ for all three values of g. If $P(CD) < P^*$ for any g, the value of n is increased by one and the process repeated until (2.3) is attained for all three values of g.

Method of Calculating n_0 for the Case k=1

We first discuss the determination of n_0 for the special case k=1. This situation has already been considered by Fhanér (1974) and Wilcox (1979). In particular, n_0 is the smallest integer n so that simultaneously

$$(3.8) \qquad \sum_{x=s}^{n} \binom{n}{x} (p_0+\delta^*)^x \, (1-p_0-\delta^*)^{n-x} \geq P^*$$

and

$$\sum_{x=0}^{s-1} \binom{n}{x} (p_0-\delta^*)^x \, (1-p_0+\delta^*)^{n-x} \geq P^*.$$

These two quantities are fairly inexpensive to evaluate on a computer, even for $n > 500$. They can also be calculated via the relationship

$$I(1,p,s,n) = \sum_{x=s}^{n} \binom{n}{x} p^x (1-p)^{n-x}$$

where $I(1,p,s,n)$ is the usual incomplete beta function. It has also been shown that an approximate value of n_0 is given by $\lambda^2 p_0(1-p_0)/(\delta^*)^2$ where λ is the P^* quantile of the standard normal distribution.

The Case k=2

For k=2 there are three values of g that need to be considered. For g=2, the minimum P(CD) is given by $I(2,p,s,n)$ with $p=p_0+\delta^*$. From Sobel et al. (1977, p.8)

$$I(2,p,s,n) = \sum_{y=2s}^{n} \binom{n}{y} p^y (1-2p)^{n-y} \Big[\sum_{z=s}^{y-s} \binom{y}{z}\Big].$$

(It might appear that the term $p^y(1-2p)^{n-y}$ should be either $(2p)^y(1-2p)^{n-y}$ or $p^y(1-p)^{n-y}$, but from Sobel et. al it can be seen that this expression is correct.) For g=0, the minimum P(CD) is given by

$$J(k,p,s,n) = \sum_{y=0}^{k} (-1)^y \binom{k}{y} I(y,p,s,n) \tag{3.9}$$

with $p=p_0-\delta^*$ where $I(0,p,s,n)=1$. In fact, from Sobel and Uppuluri (1974), it follows that for g=0, the minimum P(CD) is given by (3.9) for any k. For g=1, the minimum P(CD) is

$$\sum_{x_1=0}^{s-1} \binom{n}{x_1} (p_0-\delta^*)^{x_1} (1-p_0+\delta^*)^{n-x_1} I(1,(p_0+\delta^*)/(1-p_0+\delta^*),s,n-x_1).$$

The last expression is obtained by writing the P(CD) as is done in (3.7).

An Approximate Solution for k > 1.

For k > 2, the necessary calculations to compute n_0 become prohibitively expensive. In many cases, however, exact results are possible by first applying the approximate solution about to be described and then performing the calculations outlined below.

The proposed approximate solution is based on the Bonferroni inequality which states that for any set of events $B_1,\ldots,B_m$,

$$(3.10) \quad P(\cap B_i) \geq 1 - \Sigma P(B_i^c)$$

where B_i^c is the complement of the event B_i. Several other approximate solutions were investigated that relied on the central limit theorem and various inequalities for the multivariate normal distribution. However, the procedure proposed here is relatively easy to use, it is inexpensive, and it is surprisingly accurate.

Familiarity with the multinomial distribution suggests that when p_0 is close to zero, as is typically the case for the problem under investigation, the P(CD) is a minimum when g=k. Conditions under which this is true are not known. In all cases considered, however, it was verified that this is indeed the case. Fortunately it is possible to arrive at this conclusion for the special cases considered here without calculating the exact value of the P(CD) for every g. This point is illustrated below.

Let n_1 be the smallest integer such that (3.8) is greater than or equal to T^* where $T^*=1-(1-P^*)/k$. We consider n_1 as a first approximation to n_0. As alluded to earlier, our main motivation for using n_1 to approximate n_0 is the high cost of determing n_0 exactly for $k\geq3$. Before considering this case, it is of interest to examine the accuracy of the approximation for k=2.

Table 4 gives the value of n_1 for k=2 and the values of P^* and δ^* used in Table 2. As can be seen, n_1 gives a good approximation to n_0.

The Case k=3

The first step used to determine n_0 exactly for the case k=3 was to compute n_1 in the manner described in the previous section. The results are reported in Table 5. Next, using the value of n_1 the value of $I(3,p_0+\delta^*, s,n_1)$ was calculated. This was accomplished with the reduction formula

$$(3.11) \qquad I(k,p,s,n) = \sum_{y=(k-1)s}^{n-s} \binom{n}{y}(1-p)^y p^{n-y} I(k-1, p/(1-p),s,y)$$

given by Sobel et al. (1977, p.8). The value of n_1 was then adjusted to find the smallest value of n_1, say n_2, so that $I(k,p_0+\delta^*,s,n_2) \geq P^*$. A comparison of Table 5 with Table 3 shows that frequently $n_0=n_2$ and that typically the value of n_1 is within one of the value n_0.

Finally, to verify that n_2 is sufficiently large to satisfy (2.3), i.e., that $n_0=n_2$, we calculated $I(i,p_0+\delta^*,s,n_2)$ for i=1,2 and $J(i,p_0-\delta^*,s,n_2)$ for i=1,2,3. As previously pointed out $J(i,p_0-\delta^*,s,n_2)$ is the probability of correctly classifying the i distractors having probability $p=p_0-\delta^*$ of being chosen by a randomly selected examinee. These values were then used in conjunction with the Bonferroni inequality to show that $P(CD) \geq P^*$.

As an illustration, consider the case k=3, $p_0=.1$, $\delta^*=.05$ and $P^*=.95$. The value of n_1 was found to be 199 and it was verified via (3.11) that n=199 is the smallest sample size so that $P(CD)\geq P^*$ when g=k (all distractors have a probability of being chosen by a typical examinee that is greater than the standard p_0). Consider, for example, the case g=1. It was found that $I(1,p_0+\delta^*,s,199)=.996$, and that $J(2, p_0-\delta^*,s,199)=.995$. As explained earlier, the first quantity is the probability of making a correct decision for a distractor having $p=p_0+\delta^*$ and the second quantity is the probability of a correct decision for two distractors having $p=p_0-\delta^*$. Applying (3.10) it follows that the joint probability of correctly classifying all three distractors is greater than or equal to 1-(1-.996)-(1-.995)=.991. Thus, the

the desired probability guarantee is satisfied for this special case. Proceeding in a similar manner, it can be seen that $n=199$ is sufficiently large for $g=2$ as well.

The Case k=4

The last situation considered is $k=4$. In this case the value of n_0 was approximated in the manner previously described, but no attempt was made to make an exact evaluation of the P(CD) under the least favorable configuration. However, checks were made on the adequacy of n_0 with a normal approximation to $I(k,p,s,n)$ given by

$$A_k(\rho,h) = \Phi^k(h)+\rho\binom{k}{2}\phi^2(h)\Phi^{k-2}(h)+\frac{\rho^2}{2}\{\binom{k}{2}h^2\phi^2(h)\Phi^{k-2}(h)-6\binom{k}{3}\phi^3(h)\Phi^{k-3}(h)+6\binom{k}{4}\phi^4(h)\Phi^{k-4}(h)\}. \tag{3.12}$$

where $\rho=-s(n-s+1)^{-1}$ and $h=2\,(\text{arc sin}\ p^{\frac{1}{2}} - \text{arc sin}\ [s/(n+1)]^{\frac{1}{2}})(n+2)^{\frac{1}{2}}$, Φ is the standard normal cumulative distriubtion function and ϕ is the standard normal density function. This approximation was proposed by Sobel et al. (1977, section 2.4) who claim that it generally gives better results than the normal approximation to the discrete multinomial distribution.

Table 6 gives the resulting values of n_0 for $k=4$. Using (3.12) in conjunction with the Bonferroni inequality, an approximate lower bound to the P(CD) was also determined for each n_0. These values are reported in Table 7.

4. A Lower Bound to the P(CD) for a Typical Item.

In this section we describe how a retrospective study might be conducted to estimate a lower bound to the P(CD) for a typical item under study. Before doing so, we note that once observations are available it is also of interest to obtain a point estimate of the P(CD) for a typical item and that under certain circumstances a theoretical solution to this problem exists. For example, we might assume that $p_1,\ldots,p_k$ arise from a Dirichlet distribution the parameters of which can be estimated in the manner described by Mosimann (1962). However, there remains the practical difficulty of evaluating the P(CD) once an expression for it has been obtained. For this reason we do not discuss this problem further.

Although there are difficulties with obtaining a point estimate of the P(CD) for a typical test item, it is fairly easy to obtain a lower bound to this quantity by proceeding in the manner about to be described. It is assumed that observations are available on N items under investigation. Consider the first distractor of every item having probability $p_{1j}(j=1,\ldots,N)$ of being chosen by a typical examinee. Let $h_1(p)$ be the marginal distribution of p_{1j}. No assumption is made about the form of h; it is merely assumed that the first two moments of h exist. Assuming the conditional distribution of x_{1j} is binomial for a given p_{ij}, we can estimate the mean and variance of p_{1j} over the domain of items, say μ and σ^2, with

$$\hat{\mu} = (Nn)^{-1} \sum_j x_{1j}$$

and

$$\hat{\sigma}^2 = \hat{\mu}_1 - \hat{\mu}^2$$

where

$$\hat{\mu}_1 = [Nn(n-1)]^{-1} \sum_j (x_{1j}^2 - x_{1j})$$

(Lord and Novick, 1968, p. 521).

Henceforth we assume μ and σ^2 are known. Let

$$\begin{aligned} \xi &= p_0, \text{ if } \mu < p_0 \\ &= \mu, \text{ if } p_0 \leq \mu \leq 1 \end{aligned}$$

and

$$\begin{aligned} U &= \frac{\sigma^2}{\sigma^2 + (\xi - \mu)^2}, \text{ if } 0 < \sigma^2 \leq m \\ &= (\mu(1-\mu) - \sigma^2)/(1-p_0)p_0, \text{ otherwise} \end{aligned}$$

where

$$m = \max\{\mu(p_0 - \mu), (\mu - p_0)(1-\mu)\}.$$

Let β_1 be the probability of a false-negative decision for the first distractor of a randomly chosen item, i.e., $\beta_1 = P(x_1 \leq s, p_1 \geq p_0)$. Using results given by Skibinsky (1977), Wilcox (1979C) shows that

$$\beta_1 \leq U \sum_{x=0}^{s-1} \binom{n}{x} p_0^x (1-p_0)^{n-x}.$$

The details of the argument are given by Wilcox and so they need not be repeated here. Let $\alpha_1 = P(x_1 \geq s_1, p_1 \leq p_0)$. It can also be shown that

$$\alpha_1 \leq U_1 \sum_{x=s}^{n} \binom{n}{x} p_0^x (1-p_0)^{n-x}.$$

where U_1 is the value of U with ξ replaced with

$$\xi_1 = \mu, \text{ if } \mu < p_0$$
$$= p_0, \text{ if } p_0 < \mu \leq 1$$

Using the above procedure, we obtain an upper bound to α_i and β_i, say $\tilde{\alpha}_i$ and $\tilde{\beta}_i$, for i=1,...,k. From the Bonferroni inequality it follows that

$$P(CD) \geq 1 - \sum_i (\tilde{\alpha}_i + \tilde{\beta}_i)$$

It is also of interest to note that a lower bound to the P(CD) can be determined for a given $\delta^* > 0$. The interested reader is referred to Wilcox (1979C).

5. Comparing Two Binomial Probability Functions

As pointed out in the introduction to this paper, there may be situations where an investigator is interested in ascertaining the effect of a particular modification to a multiple-choice test item under study. It was further suggested that this problem might be formulated in terms of comparing a binomial probability function to a control. That is, there are two binomial probability functions having probability of success p' and p and the goal is to determine whether $p' < p$. A solution to this problem is given by Wilcox (1979b). Here we extend this solution to cases where we want to determine whether $p' < p \pm c$ where c is a constant specified in advance by the investigator as being appropriate for the situation at hand. In other words, we want to determine whether the difference between p' and p is reasonably large.

Let x and y be the observed number of successes corresponding to the populations having probability of success p' and p, respectively. The decision $p'<p+c$ is made if $n^{-1}x<n^{-1}y+c$; otherwise the reverse is said to be true.

As before, an indifference zone formulation of the problem is used. In this case the indifference zone consists of the open interval $(p+c-\delta^*, p+c+\delta^*)$. If $p+c-\delta^*<p'<p+c+\delta^*$ the investigator is not particularly concerned about which decision is made. If $p'<p+c-\delta^*$ or if $p' \geq p + c + \delta^*$ we want the probability of a correct decision to be reasonably high.

Since the family of binomial probability functions has the monotone likelihood ratio property, it can be seen that for fixed p, the minimum P(CD) is

$$(5.1) \quad \sum_{y=0}^{n} \sum_{x=0}^{y-1+[nc]} \binom{n}{x}(p+c-\delta^*)^x (1-p-c+\delta^*)^{n-x} \binom{n}{y} p^y (1-p)^{n-y}$$

or

$$(5.2) \quad \sum_{y=0}^{n} \sum_{x=y+[nc]}^{n} \binom{n}{x}(p+c+\delta^*)^x (1-p-c-\delta^*)^{n-x} \binom{n}{y} p^y (1-p)^{n-y},$$

whichever is smaller, where [nc] represents the largest integer less than or equal to nc. Thus, to guarantee that both (6.1) and (6.2) have a value exceeding P^*, it is sufficient to minimize these quantities as a function of p and see whether the desired condition holds for a given n. If, after minimization, either (6.1) or (6.2) is less than P^*, a larger value of n must be used. Table 8 gives the smallest required sample sizes for P^*=.75, .9, .95, .99; δ^*=.1 and c=0, .05, .1, .15.

Concluding Remarks

The main result in this paper is that a researcher can solve the following type of problem. Suppose we have a multiple choice test item with k=4 distractors. Further suppose we want to determine which distractors have a probability of less than .1 of being chosen by a typical examinee, and *simultaneously* determine which have a probability of at least .1. A decision about each distractor is made based on a random sample of examinees. If the proportion of examinees choosing a distractor is less than .1, we decide the corresponding cell probability is less than .1; otherwise the reverse decision is made. What is the minimum number of examinees required so that regardless of the actual cell probabilities, a correct sort of the distractors is made with probability *at least* .975 when an indifference zone of $\delta^*=.05$ is used? From Table 7, the answer is n=269. If instead there are k=2 distractors, Table 2 says that at least n=235 examinees would be needed.

While the original motivation for this paper was to analyze distractors, an additional application of the results reported here recently came to the author's attention. Macready and Dayton (1977) illustrate how latent structure models might be used to measure achievement. For the simplest case, we have two equivalent items for measuring a particular skill. Two items are defined to be equivalent if every examinee knows the answer to both or neither one. Let ζ be the proportion of examinees who have acquired the skill, and let β_i=P(correct on the *ith* item | examinee does not know), i=1,2. For a randomly selected examinee, the probability of a correct on the first item and an incorrect on the second is

$$P_{10} = \beta_1(1-\beta_2)(1-\zeta)$$

and the probability of incorrect and then a correct is

$$p_{01} = (1-\beta_1)\beta_2(1-\zeta).$$

If we assume $\beta_1=\beta_2<\frac{1}{2}$, then $p_{10}<\frac{1}{4}$, and $p_{01}<\frac{1}{4}$. Since p_{10} and p_{01} are cell probabilities of a multinomial distribution, a partial check on the model can be made by estimating p_{10} and p_{01} in the usual manner, and seeing whether the values are both less than ¼. Determining the number of examinees required can be accomplished with the results given in this paper.

REFERENCES

Fhanér, S. Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Gibbons, J., Olkin, I., and Sobel, M. Selecting and ordering populations: A new statistical methodology. New York: John Wiley, 1977.

Henrysson, S. Gathering, Analyzing, and using data on test items. In R. L. Thorndike (Ed.) Educational Measurement, American Council on Education, Washington, D. C., 1971.

Huang, W. Bayes approach to a problem of partitioning k normal populations. Bulletin of the Institute of Mathematics Academia Sinica 1975, 3, 87-97.

Keats, J. A. and Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lord, F. M. and Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison - Wesley, 1968.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Mosimann, J. E. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika, 1962, 49, 65-82.

Olkin, I. and Sobel, M. Integral expressions for tail probabilities of the multinomial and negative multinomial distributions. Biometrika, 1965, 52, 167-179.

Skibinsky, M. The maximum probability on an interval when the mean and variance are known. Sankhya, 1977, Series A, 39, 144-159.

Sobel, M., Uppuluri, V. R. R. and Frankowski, K. Selected tables in mathematical statistics, volume IV. Providence, Rhode Island: American Mathematical Society, 1977.

Sobel, M. and Uppuluri, V. R. R. Sparse and crowded cells and Dirichlet distributions. The Annals of Statistics, 1974, 2, 977-987.

Tong, Y. L. On partitioning a set of normal populations by their locations with respect to a control. Annals of Mathematical Statistics, 1969, 40, 1300-1324.

Wilcox, R. R. Comparing examinees to a control. Psychometrika, 1979, 44, 55-68 (a).

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement, 1979, 31, 13-22 (b).

Wilcox, R. R. On false-positive and false-negative decisions with a mastery test. Journal of Educational Statistics, 1979, 4, 59-73 (c).

TABLE 1

Values of n_0 for k=1

P_0	δ^*	P^*: .75	.9	.95	.975	.99
.1	.05	18	60	110	160	239
.15	.05	25	86	153	219	313
.15	.10	5	20	40	59	86
.2	.05	33	109	180	260	370
.2	.10	9	25	45	70	100

TABLE 2

Values of n_0 for k=2

P_0	δ^*	P^*: .75	.90	.95	.975	.99
.10	.05	49	110	160	235	290
.15	.05	66	153	219	292	380
.15	.10	19	40	59	79	106
.20	.05	84	180	260	340	455
.20	.10	24	45	70	90	120

TABLE 3

Values of n_0 for k=3

p_0	δ^*	P^*: .75	.9	.95	.995	.99
.1	.05	70	140	199	250*	320*
.15	.05	100	192	259	333*	420*
.15	.10	26	52	72	92	119
.2	.05	120	225	305	390*	495*
.2	.10	30	60	80	105	135

Entries marked with an * were not verified using exact calculations of the P(CD).

TABLE 4

Value of n_1 for k=2

p_0	δ^*	P^*: .75	.9	.95	.975	.99
.1	.05	49	110	160	235	290
.15	.05	66	153	219	292	380
.15	.10	19	40	59	79	106
.2	.05	89	180	260	345	460
.02	.10	24	45	70	90	120

TABLE 5

Values of n_1 for k=3

p_0	δ^*	P^*: .75	.9	.95	.975	.99
.1	.05	79	140	199	250	320
.15	.05	100	193	260	333	420
.15	.10	26	52	72	93	119
.2	.05	120	225	305	***	***
.2	.10	30	60	80	105	135

TABLE 6

Approximate Values of n_0 for k=4

p_0	δ^*	P^*: .75	.9	.95	.975	.99
.1	.05	99	160	219	270	349
.15	.05	132	219	292	360	446
.15	.10	33	59	79	99	126
.20	.05	155	260	345	425	525
.20	.10	39	64	85	105	135

TABLE 7

Values of n_0 for k=4 Using (3.12)

p_0	δ^*	P^*: .75	.9	.95	.975	.99
.1	.05	99	160	219	269	348
.15	.05	132	219	292	359	453
.15	.10	33	59	79	99	125
.20	.05	155	260	345	425	540

TABLE 8

Values of n for comparing a binomial distribution to a control, $\delta^*=.1$

c	P^*: .75	.9	.95	.99
0	32	91	144	245
.05	41	101	161	261
.10	41	101	151	261
.15	34	94	141	254

(A)

To appear in Applied Psychological Measurement

SOLVING MEASUREMENT PROBLEMS WITH AN ANSWER-UNTIL-CORRECT SCORING PROCEDURE

Rand R. Wilcox

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California . Los Angeles

and the

DEPARTMENT OF PSYCHOLOGY
University of Southern California

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

ABSTRACT

Answer-until-correct (AUC) tests have been with us for some time. Pressey (1950) points to their advantages in enhancing learning, and Brown (1965) has proposed a scoring procedure for it that appears to increase reliability (Gilman and Ferry, 1972; Hanna, 1975). This paper describes a new scoring procedure for AUC tests that solves various measurement problems. In particular, it makes it possible to check whether guessing is at random, it gives a measure of how "far away" guessing is from being random, it corrects observed test scores for partial information, and it yields a measure of how well an item reveals whether an examinee knows or does not know the correct response. In addition, the paper derives the optimal linear estimate (under squared error loss) of true score that is corrected for partial information, and it derives another formula score under the assumption the Dirichlet-multinomial model holds. Once certain parameters are estimated, the latter formula score makes it possible to correct for partial information using only the examinees usual number correct observed score. The importance of this formula score is discussed at the end of the paper. Finally, various statistical techniques are described that can be used to check the assumptions underlying the proposed scoring procedure.

INTRODUCTION

When an examinee responds to a multiple-choice test item, there is the problem that an examinee's response might not reflect his/her true state. The most obvious example, and the one of central concern here, is that an examinee might guess the correct response without knowing what it really is. The common solution to this problem is to assume guessing is at random. That is, if there are t alternatives from which to choose, and only one is correct, the probability of a correct response when the examinee does not know is t^{-1}. Simultaneously, however, it is recognized that to assume random guessing is indefensible. One possibility is that an examinee might be able to eliminate one or more distractors without knowing the correct response. In support of this possibility are empirical investigations on formula scoring where it was found that the probability of guessing is substantially higher than would be expected when random guessing occurs (Bliss, 1980; Cross and Frary, 1977). We might assume guessing is at random anyway, but this can have serious consequences in terms of test accuracy (e.g., Weitzman, 1970; Wilcox, 1980).

The purpose of this paper is to examine how an answer-until-correct (AUC) testing procedure might be used to take into account the effects of guessing. One advantage of the proposed scoring procedure is that its efficacy can be empirically checked in several different ways. The model contains number-right scoring, as well as the assumption of random guessing, as a special case. Thus, when observed test scores suggest that the model holds, the appropriateness of the two more common scoring procedures can be checked, as is illustrated in a later section of the paper. On a related matter, the model can be used to test whether items are "ideal" in the sense defined by Weitzman (1970). This just means that a random guessing assumption can be tested. Using the entropy function, it is also possible to measure how "close" the probability of guessing is to t^{-1}. This is important because when the probability is not close to t^{-1}, this suggests it might be possible to improve the distractors which in turn will improve test accuracy. The exact sense in which this is true is explained below. Another advantage of the model is that it yields a measure of test accuracy that is not ordinarily available. Two new formula scores are also derived, the advantages and disadvantages of which are discussed below.

It should be noted that a scoring rule for an AUC test has been proposed by Brown (1965). The scoring rule has been empirically investigated by Gilman and Ferry (1972) and Hanna (1975) who found it to be more reliable than number correct scoring. Moreover, an AUC testing procedure has been advocated from the standpoint of enhancing learning (Pressey, 1950). The goal in this paper is to propose a different scoring rule that corrects for partial information.

ASSUMPTIONS

It is assumed that when an examinee responds to an achievement test item, he/she can be described as either knowing or not knowing the correct response. In the terminology of Reulecke (1977) this means that the model includes a binary structure variable, or following Harris and Pearlman (1978) examinees are described in terms of a dichotomized latent trait. One more possibility is to say that an examinee either has or has not acquired the "psychological structure" of a task (Spada, 1977). This means that the model is deterministic in the sense that if an examinee's latent state is known, and if there are no errors at the item level, it would be known whether an examinee would produce a correct response. However, the model includes what Reulecke (1977) calls an intensity variable. In particular, it is assumed that an examinee who does not know might give a correct response. The probability of this event is unknown, but it can be estimated with the scoring formula and probability model described below.

Following Horst (1933), it is assumed that when an examinee does not know, he/she can eliminate at most t-2 distractors from consideration. Once these distractors are eliminated, the examinee chooses an answer at random from among those that remain. An examinee who knows, always gives the correct response.

Finally, an answer-until-correct scoring procedure is assumed. This means that an examinee responds to a test item until the correct alternative is chosen.

THREE TYPES OF GUESSING

Before turning to the new results, it is important to be more precise about what is meant by guessing. Three types can be described. The first applies to a situation where randomly sampled examinees respond to the same multiple-choice item. In this case we define guessing as the probability of a correct response given that the randomly sampled examinee does not know. The second, or Type II guessing, is defined in terms of a single examinee responding to an item randomly sampled from some item domain. The rate of guessing for the examinee is the probability of a correct response to a randomly sampled item that he/she does not know. Finally, there is Type III guessing which is the probability of a correct response over independent repeated trials where a single examinee responds to a specific item he/she does not know. Wilcox (1977a) examines some latent structure models that are relevant to this case, but there are some practical difficulties (Wilcox, 1979) which limit their use. Only Type I and Type II guessing are considered.

A MODEL FOR AUC TESTS AND TYPE I GUESSING

Consider a randomly sampled examinee responding to a specific test item using an AUC test. For convenience, particular attention is given to the case where the multiple-choice test item has t=4 alternatives from which to choose, one of which is correct. The results are readily extended to any value of t. Based on the above assumptions, the examinee belongs to one of t=4 mutually exclusive groups. In particular, the examinee knows the correct response, or can eliminate 0, 1, or 2 distractors. Let ζ be the proportion of examinees who know, and let ζ_i be the proportion of examinees who can eliminate i distractors. The probability of a correct response the first time a randomly selected examinee chooses an alternative is

eta l.c.

Insert Equation 1 here

The probability of an incorrect on the first choice and a correct on the second is

Insert Equation 2 here

The probability of two misses and then a correct is

Insert Equation 3 here

and the probability of three incorrects is

Insert Equation 4 here

More generally,

Insert Equation 5 here

where i=2,..., t.

For a random sample of N examinees let x_i be the number who correspond to the event associated with p_i. For example, x_1 is the number of examinees who are correct on the first alternative chosen, and x_2 is the number of examinees who are incorrect and then correct. The x_i's have a multinomial probability function given by

Insert Equation 6 here

where

Insert Equation 7 here

Since $\zeta=p_1-p_2$,

Insert Equation 8 here

is an unbiased estimate of ζ. From Zehna (1966) it also follows that $\hat{\zeta}$ is an unrestricted maximum likelihood estimator. Proceeding in a similar manner also yields unbiased, unrestricted, maximum likelihood estimates of the ζ_i's, namely,

Insert Equation 9 here

Insert Equation 10 here

Insert Equation 11 here

Note the model assumes that

Insert Equation 12 here

Maximum likelihood estimates of the ζ's are available under this restriction of the p_i's as noted by Barlow et al. (1972). For example, the maximum likelihood estimate of ζ, assuming equation 12 holds, is given by equation 8 when $x_1 \geq x_2$, and it is $\hat{\zeta}=0$ otherwise.

Using the Model to Analyze Achievement Test Items

Macready and Dayton (1977) describe a probability model based on Type I guessing that might be used to analyze mastery tests consisting of equivalent items. This section illustrates how the above model can be used to analyze achievement test items in a similar but different fashion.

Suppose, as is customary, it is decided that an examinee knows the correct response if the first alternative chosen is the correct answer, and that otherwise the examinee does not know. In this case a test constructor would like to know the accuracy of the decision about a typical examinee based on his/her response.

The cells in Table 1 give the probability of the four possible outcomes when an examinee responds to an item.

Insert Table 1 here

Thus for a randomly sampled examinee, the probability of a correct decision about an examinee's latent state is the proportion of agreement in Table 1, namely,

Insert Equation 13 here

An unrestricted maximum likelihood estimate of P is just

Insert Equation 14 here

where $\hat{\zeta}$, $\hat{\zeta}_0$, $\hat{\zeta}_1$ and $\hat{\zeta}_2$ are given by equations 8-11. For any t,

Insert Equation 15 here

P can also be estimated assuming equation 12 holds, as is illustrated below. In many instances this will yield the same estimate of P as is given by equation 14, but this is not always the case.

Using equation 13, it would seem that for any fixed ζ, the accuracy of an item is maximized when guessing is at random, i.e., when $\zeta_1=\zeta_2=0$ and $\zeta_0=1-\zeta$. This can be established in a more formal manner as follows: The inequality

Insert Equation 16 here

holds whenever $x_1 \leq x_2 \leq \ldots \leq x_N$ if and only if

Insert Equation 17 here

gma, cap and $\Sigma c_i = \Sigma b_i$ (e.g., Marshall and Olkin, 1979, p. 445). It follows that P is maximized when $\zeta_1=\zeta_2=0$ since equation 17 holds when $\underline{c}=(\zeta, \zeta_0, \zeta_1, \zeta_2)$ and $\underline{b}=(\zeta, 1-\zeta, 0, 0)$.

Another way to characterize Table 1 is to use the "del" measure developed by Hildebrand et al. (1977) which, for the situation at hand, is equivalent to Cohen's kappa (Cohen, 1960). In terms of the ζ's, this measure of association is

Insert Equation 18 here

where

Insert Equation 19 here

appa, l.c. Following Hildebrand et al., κ can be interpreted as follows: Suppose it is desired to measure the extent to which an examinee's latent state can be "predicted" according to the decision rule being used. The off-diagonal cells in Table 1 represent the error rates. The index κ represents the proportional reduction in the number of cases in the pair of error cells when a shift is made from statistical independence with the population marginals to the actual probability structure.

Note that Equation 18 is the value of κ assuming the model holds.

A Measure of Item "Idealness"

Weitzman (1970) describes an asymptotic test of whether an item is ideal. As previously indicated, an item is defined to be ideal if guessing is at random. In the above notation, this corresponds to having $\zeta_1=\zeta_2=0$ which implies that $p_2=p_3=p_4$. A practical problem is that the null hypothesis that $p_2=p_3=p_4$ might be tested and rejected, when in fact p_2, p_3 and p_4 are nearly the same in value. This in turn might lead to efforts in improving the distractors when the item is already close to being ideal.

The simplest approach to this problem is to estimate ζ_1 and ζ_2 and see how close they are to zero. If they are not, simply examine the distractors and decide whether any of them can be improved. Some additional possibilities are described and illustrated below.

When trying to determine whether ζ_1 and ζ_2 are both close to zero, it might be desirable to take into account their combined effect on how close the item is to being ideal. Looking at ζ_1 and ζ_2 separately, they might appear to be close to zero, but together, perhaps the item could be improved by a substantial amount. The problem becomes more complex when more than three distractors are used. Thus, it would be convenient to have some measure of how well an item approximates the ideal situation where $\zeta_0=1-\zeta$.

One approach is to estimate ζ which yields an estimate of the proportion of agreement in Table 1 for the case $\zeta_0=1-\zeta$. Thus, we have estimated the maximum possible value of P for fixed ζ, say P_{max}, which corresponds to the estimated value of ζ. For t=4, $P_{max}=\frac{3}{4}+\frac{\zeta}{4}$. Next, estimate P which yields an estimate of

Insert Equation 20 here

This gives a measure of how ideal the item really is. When the model
Delta, cap holds, $\Delta \geq 0$, and the closer Δ is to zero, the better the item.

Employing the Δ measure seems to be intuitively appealing, and in some situations it might suffice. However, there are at least two objections to its use. First, it has been suggested (e.g., Marshall and

Olkin, 1979, p. 408) that measures of inequality should have certain properties, namely, they should be Schur-convex, or strictly Schur-convex. Here the goal is to measure the inequality of p_2, p_3, p_4. (The meaning of a Schur-convex function is not given since it does not play a direct role in the results to follow. The interested reader is referred to Marshall and Olkin, Chapter 3.) This requirement was first formulated by Dalton (1920), and steps in this direction were taken by Lorenz (1905) and Pigou (1912). Thus, as a measure of the inequality of p_2, p_3 and p_4, Δ might be objectionable because it is not Schur-Convex. To see this, it is sufficient to observe that Δ, as a function of p_2, p_3 and p_4, is not symmetric. The second objection is that even when the model holds, the estimate of ζ_1 and ζ_2 can be negative, and the estimate of ζ_0 can be greater than one. In this case Δ cannot be interpreted as a difference of two probabilities. Perhaps we could use Δ anyway, but an investigator might prefer to use a more traditional index of inequality.

For the problem at hand, the index of inequality that suggests itself is the entropy function. The entropy of a probability mass function $p_k \geq 0$, $k=1, \ldots, r$, is

Insert Equation 21 here

where $\Sigma p_k=1$. (In some instances, the logarithms in equation 21 are taken to the base 10 or the base 2. See Kullback, 1959, p. 7.) The function H provides a measure of the degree of uniformness of a distribution. That is, the larger is H, the more uniform is the distribution. The minimum value of H occurs when $p_1=1$, its maximum value occurs when $p_1=\ldots p_r=1/r$, and it is Schur-concave (implying that -H is Schur-convex). See Marshal and Olkin (1979, chapter 13, section E). To measure the idealness of an item, the inequality of p_2, p_3 and p_4 needs to be measured which suggests that $H(q_1, q_2, q_3)$ be used where $q_i=p_{i+1}/(1-p_1)$, $i=1,2,3$. In this case the maximum possible value of H occurs when $q_i=(t-1)^{-1}$. An additional reason for using the entropy function is given in the next section of the paper. Brown (1965, section 3) also used the entropy function but in a slightly different fashion.

Empirical Checks on the Model

From equations 1-4 various restrictions on the p_i's are evident in order for the model to hold. For instance, it requires having $p_1 \geq p_2 \geq p_3 \geq p_4$. This assumption can be tested using results reported by Robertson (1978). It should be noted that when $p_1=p_2$, the probability of having $x_1 \geq x_2$ approaches .5 as N, the number of examinees, gets large. Thus, there is a reasonably high probability that the usual estimate of the p_i's will indicate that the model does not hold when the p_i's are approximately equal in value. Of course, the hypothesis $p_2=p_3=p_4$ can be tested, but this does not give a direct measure of how ideal an item is. The null hypothesis might be rejected, for example, but this does not directly indicate the extent to which p_2, p_3 and p_4 are unequal. Another approach might be to estimate H, especially when the data suggests the model might not hold, and if H is reasonably close to its maximum value, decide that the item is ideal. We are not suggesting that hypothesis testing be discarded all together, the point is that the entropy function gives us some additional information about how close an item is to being ideal that is otherwise unavailable. It might help to note that a similar situation occurs in the analysis of variance (Hays, 1973, pp. 484-488).

Another requirement of the model is that $p_4 \leq \frac{1}{4}$, otherwise, $\zeta_0 > 1$. For similar reasons the model requires that $p_3 - p_4 \leq 1/3$ and $p_2 - p_3 \leq \frac{1}{2}$. However, $p_1 \geq p_2 \geq p_3 \geq p_4$ implies that these additional inequalities are true.

Illustrations

The results given above are illustrated with test scores for students enrolled in an undergraduate psychology course at the University of Southern California. Each item had t=5 distractors. There were four test forms, and each form had forty items. For simplicity, only 4 items are analyzed, and only one test form is used. A more extensive analysis of the data, together with some new theoretical results, will appear in a forthcoming report.

Table 2 gives the observed frequencies of the number of examinees who got the item correct on the ith attempt (i=1,...,5). For example, there were 42 examinees who were incorrect on their first attempt, but were correct on their second attempt of item 2.

Insert Table 2 here

The first step when applying the results given above is to test the hypothesis that equation 12 holds. As previously mentioned, this is accomplished with results in Robertson (1978). This was done for all 40 items on the test using a .01 level of significance. For items 1 and 2 in Table 2, applying Robertson's test is not necessary since the estimate of the p_i's already satisfies equation 12. Item 3 is highly nonsignificant, but the null hypothesis is rejected for item 4.

For 21 of the 40 items, Robertson's test was unnecessary since the estimate of the p_i's satisfied equation 12. For the remaining items, the null hypothesis was rejected only once; this was for item 4 in Table 2.

Next suppose a test constructor wants to determine whether a conventional scoring procedure will yield reasonably accurate decisions about whether an examinee has acquired the skills represented by items 1, 2 and 3 in Table 2. An estimate of P via equation 15 yields a partial solution to this problem. For items 1 and 2, the estimate of ζ is $(139-14)/168=.744$ and $(100-42)/168=.345$, respectively. Thus, the corresponding estimates of P are .917 and .75.

As for item 3, estimating p_3 and p_4 under the assumption that equation 12 holds requires an application of the pool-adjacent-violators algorithm in Barlow et al. (1972, pp. 13-18). The result is $\hat{p}_3=\hat{p}_4=(29+16)/(2(168))=.134$. The estimate of ζ is .202, and so the estimate of P is .797. Note that using the pool-adjacent violators algorithm yields the same estimate of P as is obtained when equation 15 is used and when p_i is estimated with x_i/n. However, when $x_1<x_2$, using $\hat{p}_i=x_i/N$ will yield different results. The reason is that the maximum likelihood of ζ, assuming equation 12, is $\hat{\zeta}=0$ when $x_1<x_2$, and it is $(x_1-x_2)/N$ otherwise. Consider, for example, item 4 in Table 2. $\hat{\zeta}=0$, and the maximum likelihood estimate of p_2, assuming equation 12, is .369. Thus, the estimate of P is .63. If, however, we use $\hat{p}_i=x_i/N$, the estimate of P is .446.

Suppose the first three items in Table 2 constituted the whole test. Another important point is that the estimates of P yield an estimate of γ, the expected number of correct decisions for the n items on the test. The estimate is simply the sum of the estimated P values. For the case at hand γ is estimated to be 2.46. Thus, when a conventional scoring

mma, 1.c.

procedure is used to determine whether an examinee knows the correct response to an item, the expected number of correct decisions for the first three items in Table 2 is estimated to be 2.46.

If any of the P values is small, one possible way to improve the item is to improve the distractors. For example, efforts might be made to improve the least frequently chosen distractor.

To measure the effectiveness of the distractors, the entropy function is applied. For item 1 in Table 2, q_1=.483, q_2=.31, q_3=.138 and q_4=.069. Substituting these values into equation 21 yields H=1.172. The maximum possible value of H occurs when q_i=.25 (i=1,2,3,4) in which case H=1.386. For item 2, H=.99 and for item 3 H=1.347. Thus, the test scores indicate that the item with the most effective distractors is item 3 followed by item 1. The distractors for item 2 are the least effective having achieved 71.4% of the maximum possible entropy.

It should be pointed out that the above estimate of H for item 3 was not made under the assumption that equation 12 holds. If equation 12 is assumed, and the pool-adjacent-violators algorithm is applied, this yields $\hat{p}_1$=.405, $\hat{p}_2=\hat{p}_3=\hat{p}_4$=.1568 and $\hat{p}_5$=.125 in which case H=1.382. In either case, item 3 has the most effective distractors.

A MODEL FOR TYPE II GUESSING

In many instances a test consists of items representing skills that are thought to be most important. Moreover, there are situations where the skills on a test are the only ones that are of interest to the test constructor. However, in other situations (see, e.g., Hambleton et al., 1978) the items on a test are intended to be a representative sample of some larger item domain. The goal is to use test results to make inferences about what an examinee knows relative to the item pool. In either case, the results in the previous section are of interest. This section considers how an AUC test might be used to solve certain measurement problems when generalizing results for a single examinee to an item domain.

i, 1.c. For a specific examinee, let ξ be the proportion of skills among a domain of skills that he/she has acquired. Further suppose that each skill is represented by a multiple-choice test item having t alternatives from which to choose. Again for convenience, emphasis is given to the special case t=4. Let ξ_i (i=0, ... , t-2) be the proportion of items for which the examinee does not know and can eliminate i distractors. Once i distractors are eliminated, the examinee is assumed to guess at random from among those that remain. Let r_i be the probability of a correct on the ith attempt. Then for t=4,

Insert Equation 22 here

Insert Equation 23 here

Insert Equation 24 here

Insert Equation 25 here

If for a random sample of n items, y_i is the number of items the examinee is correct on the ith alternative chosen. An unbiased estimate of the ξ_i's can be derived just as unbiased estimates of ζ_i's were derived in the previous model. In particular, an unbiased estimate of ξ is

Insert Equation 26 here

Equation 26 is an estimate of true score that is corrected for an examinee's partial information. Note that equation 26 contains the usual correction for guessing formula score as a special case.

The Optimal Linear Estimator of ξ

Let z be a random variable that is an unbiased estimate of the unknown
T l.c. parameter θ. Under squared error loss, Griffin and Krutchkoff (1971) show that the optimal linear estimator of θ is

Insert Equation 27 here

Alpha, l.c.
Delta, l.c. where $\alpha = Var(\theta)/Var(z)$ and $\delta = (1-\alpha)E(\theta)$. In mental test theory, equation 27 is known as Kelley's linear regression estimate of true score (Kelley, 1947, p. 409). The point made by Griffin and Krutchkoff is that if an unbiased estimate of an examinee's true score is used, equation 27 is optimal regardless of the shape of true score distribution. Wilcox (1978) compares equation 27 to several other estimators assuming the binomial error model holds but where observed scores are generated according to a two-term approximation to the compound binomial error model. The results suggest that when simultaneously estimating the true score of several

examinees, the Griffin-Krutchkoff estimator should be used when an ensemble squared error loss function is being used. Furthermore, the results suggest that Kelley's linear regression estimate of ξ be employed.

It is assumed that the y_i's have a multinomial distribution and that observed test scores for N examinees are available. An estimate of $E(\xi)$, $Var(\xi)$ and $Var(y_1-y_2)$ is needed to apply results in Griffin and Krutchkoff where the expectations defining these quantities are over the population of examinees.

Let

Insert Equation 28 here

Insert Equation 29 here

where i=1, 2. Then

Insert Equation 30 here

Insert Equation 31 here

Since $cov(y_1, y_2|p_1, p_2) = -2np_1p_2$, it follows that

Insert Equation 32 here

Thus,

Insert Equation 33 here

and

Insert Equation 34 here

Letting y_{ij} and v_{ij} be the value of y_i and v_i, respectively, for the jth randomly sampled examinee, the above results suggest that $E(\xi)$ be estimated with

Insert Equation 35 here

and $E(\xi^2)$ with

Insert Equation 36 here

Thus, an estimate of $Var(\xi)$ is

Insert Equation 37 here

The variance of the marginal distribution of observed scores $(y_1-y_2)/n$ can be estimated in the usual manner, and so an estimate of the optimal linear estimator of ξ is obtained by substituting the results in equation 27. Of course, the results just given contain, as a special case, the optimal linear estimator under the assumption guessing is at random.

Numerical Illustration

As a simple illustration, suppose we have five examinees with observed y values as shown in the first two rows of Table 3, where the test length is n=10.

Insert Table 3 here

u, Tau,
igma, l.c.

Then $\hat{\mu}_\xi=.42$, $\hat{\tau}_\xi=.2$, and so $\hat{\sigma}_\xi=.0236$. The estimate of $\mathrm{var}((y_1-y_2)/n)$ is .0687. Therefore, the estimate of α is $\hat{\alpha}=.3435$, and so the estiamte of the optimal linear estimator is

Insert Equation 38 here

The value of $\tilde{\xi}$ for the five examinees are given in the last row of Table 2.

Before continuing, some additional comments about the above results are in order. First, the estimate of $\mathrm{var}(\xi)$ can be negative in which case $\hat{\alpha}=1$ is used. The same phenomenon occurs in the case considered by Griffin and Krutchkoff. Second, the optimal linear estimator of ξ derived above does not assume the model holds. It is the optimal linear estimator of p_1-p_2, but no insistence is made that $p_1 \geq p_2$. If the model holds, implying that $p_1 \geq p_2$, equation 33 is no longer true, and so the condition of having an unbiased estimate of ξ, as is assumed by Griffin and Krutchkoff, is no longer satisfied. For further comments on this approach to estimation, see Griffin and Krutchkoff (1971) and Wilcox (1978).

A Strong True Score Model

This section assumes that for any examinee, y_1 and y_2 have a multinomial probability function given by

Insert Equation 39 here

where, as before, $\xi = p_1 - p_2$ and $0 \leq \xi \leq 1$ is assumed. Equation 39 can be justified under an item sampling model, or it might give a good approximation to the joint probability function of y_1 and y_2. It should be noted that equation 39 implies that y_1 has a binomial probability function, and so when every examinee takes the same n items, the items have the same level of difficulty (Lord and Novick, 1968, chapter 23). On theoretical grounds, this implication of equation 34 is unjustifiable. However, for certain measurement problems, it appears that this might not be a serious restriction. (Wilcox, 1977, 1978; Algina and Noe, 1978). See also Subkoviak (1978).

Strong true-score models attempt to extend assumptions such as equation 39 to a population of examinees. The basic problem here is to find a family of distributions that approximates $g(\xi, p_2)$, the joint density of ξ and p_2. Once this is done, various measurement problems can be solved (e.g., Lord, 1965; Huynh, 1976; Wilcox, 1977).

Past experience with this type of problem (Keats and Lord, 1962; Lord, 1965; Wilcox, 1979) suggests approximating $g(\xi,p_2)$ with a bivariate Dirichlet function given by

Insert Equation 40 here

mma, cap

where Γ is the usual gamma function, $v_i>0$ $(i=1,2,3)$ are unknown parameters and $0\le\xi+p_2\le1$. (Marshall and Olkin, 1979, pp. 306-307 describe two other distributions to which the name "Dirichlet" is attached. Here, only equation 40 is considered.)

To estimate the v_i, proceed as follows: first, observe that the marginal distribution of ξ is beta with parameters v_1 and v_2+v_3 (e.g., Wilks, 1962). It follows that

Insert Equation 41 here

where, as before, μ_ξ is the mean of ξ over the population of examinees. For similar reasons,

Insert Equation 42 here

where μ_p is the mean of p_2. It is also known (e.g., Wilcox, 1977) that

Insert Equation 43 here

where

Insert Equation 44 here

Summarizing these resulte in matrix notation yields

Insert Equation 45 here

As previously indicated, μ_ξ and μ_ξ^2 can be estimated, which yields an estimate of s. An estimate of p_2 is $N^{-1}\Sigma y_{2j}$, and so equation 44 yields an estimate of the v_i's.

Mosimann (1962) applies the Dirichlet-multinomial model to two real data sets, he discusses how to check the implications of the model, and he gives several other results that have practical value, and so these issues are not discussed further. Since the Dirichlet-multinomial model is the multivariate analog of the beta-binomial model, additional insights into the appropriateness of the model are available from Wilcox (1981). The point is that the Dirichlet-multinomial model can be applied to AUC scoring procedures and so solve various measurement problems as previously indicated. An advantage of the model is that it allows guessing to vary over the population of examinees.

An important point is that if the model is assumed to hold, and in particular $0 \le \xi \le 1$, this suggests estimating ξ to be zero even when $\hat{\xi}<0$. In this case the estimates of $E(\xi)$ and $E(\xi^2)$ are not justified for the reasons given above, but they are still appropriate for the reasons given by Wilcox (1979).

One point that deserves special mention is that a new formula score can be derived that corrects for partial information. The derivation is essentially the same as the derivation of equation 4 in Wilcox (1979). Thus, we merely note that

Insert Equation 46 here

where B is the usual beta function. Thus, once the v_i's are estimated, we only need y_1 to estimate ξ.

DISCUSSION

One objection to the assumptions that were made is that the resulting model is too simple. For instance, it does not allow for the possibility of knowing and being incorrect, or the possibility of having misinformation. Brown and Burton (1979) describe a real situation where the latter problem occurs. Frary (1980) gives an interesting account of how misinformation can affect various scoring procedures, and Wilcox (1980) indicates the seriousness of the former problem when determining the length of a criterion-referenced test. Although the present model does not correct these problems, empirical checks on the appropriateness of the model can be made. It should be mentioned that models have been proposed for handling the two errors just described (e.g., Duncan, 1974; Macready and Dayton, 1977; Dayton and Macready, 1976). However, these models require additional assumptions that might not be met. The Macready-Dayton model, for example, assumes that equivalent items are available for measuring a particular skill. The assumption of equivalent items can be checked using a goodness of fit test (Macready and Dayton, 1977), using a procedure described by Hartke (1978), and results reported by Baker and Hubert (1977) might also be useful in this endeavor. (See, also, Wilcox, in press, a.) Here it is assumed that empirical investigations fail to support the existence of equivalent items, or that it is decided a priori that equivalent items do not exist. Finally, the Duncan model corrects for misinformation, but it assumes guessing is at random. The goal here is to avoid this restriction, or to find ways in which it can be empirically checked.

Another possible objection to the model is that it characterizes examinees as belonging to one of two mutually exclusive classes, namely, "knowing" and "not knowing." The relative merits of this approach are discussed in a more general context by Reulecke (1977), Hilke et al. (1977), Scandura (1971, 1973), and Spada (1976).

In some situations, the scoring procedure for Type II guessing might be objectionable because it penalizes an examinee for having partial information. That is, if an examinee wants to maximize his/her score (the estimate of ξ) the strategy would be to minimize y_2. This could be done by choosing an answer, and if it is wrong, deliberately choosing another response that is believed to be incorrect. In this case the examinee is not behaving in the manner assumed, and so the model is inappropriate. One approach to this problem is to have an examinee always mark his/her first and second choice without revealing which response is correct. Letting y_1 be the number of times the examinee's first choice is correct, letting y_2 be the number of times the second choice is correct, ξ is again estimated with $(y_1-y_2)/n$. Indeed, all of the previous results still hold. However, this might not eliminate the problem under discussion. Suppose, for example, that an examinee can eliminate all but two of the alternatives from consideration for every item on the test. If an examinee's two choices correspond to these two alternatives, the expected estimate of ξ is 0. However, if an examinee's first choice is between the two alternatives that contains the correct response, and if the examinee is deliberately incorrect on the second choice, the expected estimate of ξ is .5. One way to minimize this problem is to subject the items to an analysis that attempts to ensure guessing is at random. It was already indicated how this might be done. Another solution is to apply the Dirichlet-multinomial model. If estimates of the v_i's can be made available, the information on the examinee's first choice, the value of y_1, is all that is needed in order to estimate ξ. Several other strong ture-score models are currently being investigated that might be useful when addressing this problem. Another possibility is to check the assumptions of the model; if they do not hold, simply score the test using traditional techniques.

For practical purposes, perhaps the problem just described will be inconsequential; this remains to be seen. Also note that this problem is irrelevant in terms of the results given under Type I guessing.

In practice, the scoring rule proposed by Brown (1965) results in scoring t-i points when the correct response is chosen on the ith attempt of an item, where, as before, t is the number of alternatives from which to choose (e.g., Frary, 1980). Thus, the sooner an examinee identifies the right answer, the higher will be his/her score. In some cases, however, this scoring procedure is also inadequate. First, it gives credit to an examinee when a test constructor unintentionally produces ineffective distractors. Second, and perhaps most importantly, it gives a measure of partial information, but it does not tell us what an examinee knows in the sense of estimating ξ. The same is true of the other scoring procedures cited by Frary (1980), the scoring rule proposed by Coombs et al. (1956), as well as the subset selection rule proposed by Gibbons, Olkin and Sobel (1977, 1979). No claim is made that these procedures be abandoned, but as argued by Morrison and Brockway (1979), estimating ξ can be important.

Another point is that only two responses to each item are needed in order to estimate ξ for each examinee. The additional responses are needed only for checking the appropriateness of the model, and in particular, justifying $(y_1-y_2)/n$ as an estimate of ξ. In some cases n will be too small to accurately test the model. Determining whether this is the case can be accomplished with the statistical techniques described under Type I guessing.

Finally, it was suggested that the Dirichlet-multinomial distribution be considered when trying to find a strong true-score model that fits the data. It should be stressed, however, that considerably more experience with this distribution is needed before it is routinely applied. Wilcox (in press, b) got good results with the distribution using real data, but the extent to which it gives a good fit to mental test data is not known. An empirical investigation is currently underway in an attempt to partially resolve this problem. Consideration will also be given to several other strong true-score models. The results should be available in the near future.

REFERENCES

Algina, J., & Noe, M. G. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 1978, 15, 101-110.

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. Journal of Educational Statistics, 1977, 2, 217-233.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Brunk, H. D. Statistical inference under order restrictions. New York: Wiley, 1972.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Brown, J. Multiple response evaluation of discrimination. The British Journal of Mathematical and Statistical Psychology, 1965, 18, 125-137.

Brown, J. C., & Burton, R. R. Diagnostic models in basic mathematical skills, In the National Institute of Education, Testing, Teaching, and Learning: Report of a Conference on Research on Testing. Washington, D.C.: U.S. Department of Health, Educational and Welfare, 1979.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dalton, H. The measurement of the inequality of incomes. Econom. J. 1920, 30, 348-361.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Duncan, G. T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model, Journal of the American Statistical Association, 1974, 69, 50-57.

Fhanér, S. Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Frary, R. B. The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. Applied Psychological Measurement, 1980, 4, 79-90.

Gibbons, J., Olkin, I., & Sobel, M. Selecting and ordering populations: A new statistical methodology. New York: Wiley, 1977.

Gibbons, J. D., Olkin, I., & Sobel, M. A subset selection technique for scoring items on a multiple choice test. Psychometrika, 1979, 44, 259-270.

Gilman, D. A., & Ferry, P. Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 1972, 9, 205-207.

Griffin, B. S., & Krutchkoff, R. G. Optimal linear estimators: An empirical Bayes version with application to the binomial distribution. Biometrika, 1971, 58, 195-201.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hanna, G. S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. Journal of Educational Measurement, 1975, 12, 175-178.

Harris, C. W., & Pearlman, A. P. An index for a domain of completion or short answer items. Journal of Educational Statistics, 1978, 3, 285-304.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 1978, 15, 43-47.

Hays, W. Statistics for the Social Sciences. New York: Holt, Rinehart and Winston, 1973.

Hildebrand, D. K., Laing, J. D., & Rosenthal, H. Predication analysis of cross classifications. New York: Wiley, 1977.

Hilke, R., Kempf, W. F., & Scandura, J. M. Deterministic and probabilistic theorizing in structural learning. In H. Spada and F. Kempf (Eds.) Structural Models of Thinking and Learning. Bern: Haus Huber, 1977.

Horst, P. The difficulty of a multiple choice test item. Journal of Educational Psychology, 1933, 24, 229-232.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Kelley, T. L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947.

Kullback, S. Information theory and statistics. New York: Wiley, 1959.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lorenz, M. O. Methods of measuring concentration of wealth. Journal of the American Statistical Association, 1905, 9, 209-219.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marshall, A. W., & Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

Morrison, D. G., & Brockway, G. A modified beta-binomial model with applications to multiple choice and taste tests. Psychometrika, 1979, 44, 427-442.

Mosimann, J. E. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika, 1962, 49, 65-82.

Pigou, A. C. Wealth and welfare. New York: Macmillan, 1912.

Pressey, S. L. Development and appraisal of devices providing immediate automatic scoring of objective tests and concomitant self-instruction. The Journal of Psychology, 1950, 29, 419-447.

Rao, C. R. Linear statistical inference and its application. New York: Wiley, 1973.

Reulecke, W. A. A statistical analysis of deterministic theories. In H. Spada and F. Kempf (Eds.) Structural Models of Thinking and Learning. Bern: Haus Huber, 1977.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Scandura, J. M. Deterministic theorizing in structural learning. Journal of Structural Learning, 1971, 3, 21-53.

Scandura, J. M. Structural learning: Theory and research. New York: Gordon and Breach, 1973.

Spada, H. Logistic models of learning and thought. In H. Spada & F. Kempf (Eds.) Structural Models of Thinking and Learning. Bern: Hans Huber, 1977.

Subkoviak, M. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.

Weitzman, R. A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. Pearlman, & R. Wilcox Achievement Test Items - Methods of Study: CSE Monograph No. 6, Los Angeles: Center for the Study of Evaluation, University of California, 1977. (a)

Wilcox, R. R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics, 1977, 2, 289-307. (b)

Wilcox, R. R. Estimating true score in the compound binomial error model. Psychometrika, 1978, 43, 245-258.

Wilcox, R. R. Achievement tests and latent structure models. British Journal of Mathematical and Statistical Psychology, 1979, 32, 61-71.

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement, 1980, 4, Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, to appear.

Wilcox, R. R. Analyzing the distractors of multiple-choice test items or partitioning multinomial cell probabilities with respect to a standard. Educational and Psychological Measurement, in press. (a)

Wilcox, R. R. The single administration estimate of the proportion of agreement of a proficiency test scored with a latent structure model. Educational and Psychological Measurement, in press. (b)

Wilks, S. S. Mathematical statistics. New York: Wiley, 1962.

Zehna, P. W. Invariance of maximum likelihood estimation. Annals of Mathematical Statistics, 1966, 37, 744.

EQUATIONS

$$P_1 = \zeta + \zeta_0/4 + \zeta_1/3 + \zeta_2/2 \qquad [1]$$

$$P_2 = \zeta_0/4 + \zeta_1/3 + \zeta_2/2 \qquad [2]$$

$$P_3 = \zeta_0/4 + \zeta_1/3 \qquad [3]$$

$$P_4 = \zeta_0/4 \qquad [4]$$

$$P_i = \sum_{j=0}^{t-i} \zeta_j/(t-j), \qquad [5]$$

$$\binom{N}{x} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4} \qquad [6]$$

$$\binom{N}{x} = N!/(x_1!x_2!x_3!x_4!),\ x_4 = N-x_1-x_2-x_3,\ \Sigma p_i=1. \qquad [7]$$

$$\hat{\zeta} = (x_1-x_2)/N \qquad [8]$$

$$\hat{\zeta}_0 = 4x_4/N \qquad [9]$$

$$\hat{\zeta}_1 = 3(x_3-x_4)/N \qquad [10]$$

$$\hat{\zeta}_2 = 2(x_2-x_3)/N \qquad [11]$$

$$P_1 \geq P_2 \geq P_3 \geq P_4. \qquad [12]$$

$$P = \zeta + 3\zeta_0/4 + 2\zeta_1/3 + \zeta_2/2 \qquad [13]$$

$$\hat{P} = \zeta + 3\hat{\zeta}_0/4 + 2\hat{\zeta}_1/3 + \hat{\zeta}_2/2 \qquad [14]$$

$$P = \zeta + \sum_{i=2}^{t} p_i. \qquad [15]$$

$$\Sigma b_i x_i \leq \Sigma c_i x_i \qquad [16]$$

$$\sum_{i=1}^{k} c_i \geq \sum_{i=1}^{k} b_i,\ k=1, \ldots , n-1 \qquad [17]$$

$$\kappa = 1-(1-P)/B \qquad [18]$$

$$B = \left(\frac{3\zeta_0}{4} + \frac{2\zeta_1}{3} + \frac{\zeta_2}{2}\right) + (\zeta_0+\zeta_1+\zeta_2)(\zeta + \frac{\zeta_0}{4} + \frac{\zeta_1}{3} + \frac{\zeta_2}{2}). \qquad [19]$$

$$\Delta = P_{max} - P \qquad [20]$$

$$H(p_1, \ldots , p_r) = -\Sigma p_k \log_e p_k \qquad [21]$$

$$r_1 = \xi + \xi_0/4 + \xi_1/3 + \xi_2/2 \qquad [22]$$

$$r_2 = \xi_0/4 + \xi_1/3 + \xi_2/2 \qquad [23]$$

$$r_3 = \xi_0/4 + \xi_1/3 \qquad [24]$$

$$r_4 = \xi_0/4 \qquad [25]$$

$$\hat{\xi} = (y_1-y_2)/n \qquad [26]$$

$$\hat{\theta} = \alpha z + \delta \qquad [27]$$

$$v_i = y_i/n \qquad [28]$$

$$w_i = \frac{y_i}{n} \quad \frac{y_i-1}{n-1} \qquad [29]$$

$$E(v_i | p_1, p_2) = p_i \qquad [30]$$

$$E(w_i | p_1, p_2) = p_i^2 \qquad [31]$$

$$E(y_1 y_2 | p_1, p_2) = n(n-2) p_1 p_2 \qquad [32]$$

$$E(v_1-v_2) = E(\xi) \qquad [33]$$

$$E(w_1+w_2-[2y_1y_2/(n(n-2))]) = E(\xi^2) \qquad [34]$$

$$\hat{\mu}_\xi = N^{-1} \sum_{j=1}^{N} (v_{1j}-v_{2j}) \qquad [35]$$

$$\hat{\tau}_\xi = N^{-1} \sum_{j=1}^{N} w_{1j}+w_{2j}-n^{-1}(n-2)^{-1}\, 2y_{1j}y_{2j} \qquad [36]$$

$$\hat{\sigma}^2_\xi = \hat{\tau}_\xi - (\hat{\mu}_\xi)^2 \qquad [37]$$

$$\hat{\xi} = .3435(y_1-y_2)/n + .276 \qquad [38]$$

$$f(y_1,\, y_2 \mid \xi,\, p_2) = \frac{n!(\xi+p_2)^{y_1}\, p_2^{y_2}\, (1-\xi-p_2)^{n-y_1-y_2}}{y_1!y_2!(n-y_1-y_2)!} \qquad [39]$$

$$\frac{\Gamma(v_1+v_2+v_3)}{\Gamma(v_1)\Gamma(v_2)\Gamma(v_3)}\, \xi^{v_1}\, p_2^{v_2}\, (1-\xi-p_2)^{v_3} \qquad [40]$$

$$v_1(1-\mu_\xi{}^{-1})+v_2+v_3 = 0 \qquad [41]$$

$$v_1 + v_2(1-\mu_p{}^{-1}) + v_3 = 0 \qquad [42]$$

$$s = v_2 + v_3 \qquad [43]$$

$$s = \mu_\xi(1-\mu_\xi)^2\, \sigma_\xi^{-2} + \mu_\xi - 1. \qquad [44]$$

$$\begin{bmatrix} 1-\mu_s^{-1} & 1 & 1 \\ 1 & 1-\mu_p^{-1} & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ s \end{bmatrix} \qquad [45]$$

$$E(\xi|y_1) = \left[f(y_1)B(v_1,v_2,v_3)\right]^{-1} \binom{n}{y_1} \sum_{w=0}^{y_1} B(w+v_2,\ n-y_1+v_3)\ B(y_1-w+v_1+1,\ n-y_1+w+v_2+v_3) \qquad [46]$$

TABLE 1

Four Possible Outcomes When an Examinee Attempts an Item

	Decision		
Latent State	Knows	Doesn't Know	Marginal Probabilities
Knows	ζ	0	ζ
Doesn't Know	$\zeta_0/4 + \zeta_1/3 + \zeta_2/2$	$3\zeta_0/4 + 2\zeta_1/3 + \zeta_2/2$	$\zeta_0 + \zeta_1 + \zeta_2$

TABLE 2

Number of Examinees who are Correct on the ith Attempt of the Item

	Attempt				
Item	1	2	3	4	5
1	139	14	9	4	2
2	100	42	17	6	3
3	68	34	16	29	21
4	31	93	20	15	9

TABLE 3

Values of y_1, y_2 and $\tilde{\xi}$ with n=10 and N=5

y_1	5	7	6	9	2
y_2	3	1	2	0	2
$\tilde{\xi}$	.34	.48	.41	.59	.28

ACKNOWLEDGEMENTS

The author would like to thank Dr. Scott Fraser for generously supplying the data used in this study, and to thank Dr. Joan Murray for helpful comments on an earlier draft of this paper.

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Author's Address

Rand R. Wilcox
Department of Psychology
University of Southern California
Los Angeles, CA 90007

ACKNOWLEDGEMENTS

The author would like to thank Dr. Scott Fraser for generously supplying the data used in this study, and to thank Dr. Joan Murray for helpful comments on an earlier draft of this paper.

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

Author's Address

Rand R. Wilcox
Department of Psychology
University of Southern California
Los Angeles, CA 90007

A POLARIZATION TEST FOR MAKING INFERENCES ABOUT THE ENTROPY OF MULTIPLE-CHOICE TEST ITEMS

Rand R. Wilcox

DEPARTMENT OF PSYCHOLOGY
University of Southern California
Los Angeles, California 90007

and the

CENTER FOR THE STUDY OF EVALUATION
Graduate School of Education
University of California, Los Angeles 90024

ABSTRACT

Under an answer-until-correct scoring procedure, the entropy function can be used to measure the effectiveness of the distractors of a multiple-choice test item. This brief note indicates how a polarization test can be used to determine whether the entropy is large or small. Included as a special case is an exact test or whether guessing is at random.

1. INTRODUCTION

Consider a specific multiple-choice test item having k alternatives from which to choose, only one of which is the correct response. Suppose that a randomly sampled examinee responds to the item according to an answer-until-correct scoring procedure. This means that the examinee chooses alternatives until the correct response is identified. This is usually accomplished by having the examinee erase a shield on an answer sheet. The examinee knows immediately whether the correct response was chosen. If it was not, another shield is erased, and this continues until the correct response is identified. Wilcox (1981a) describes several measurement problems that this scoring procedure can solve. They include correcting for guessing without assuming guessing is at random, testing whether guessing is at random, measuring the effectiveness of distractors, and estimating the probability of correctly determining whether an examinee knows the correct response when a conventional scoring procedure is used. This last probability makes it possible to characterize n-item tests, and a relevant statistical procedure has been developed (Wilcox, in press). More recently the results in Wilcox (1981a) were extended to a strong true score model that allows guessing to vary over the population of examinees but which does not assume true score and guessing are independent (Wilcox, 1981b).

Suppose an answer-until-correct scoring procedure is used, and let q_i be the probability that a randomly selected examinee chooses the correct response on the ith attempt of the item. Wilcox (1981a) makes certain assumptions about how examinees behave when attempting a multiple-choice item which imply that

$$q_1 \geq q_2 \geq \cdots \geq q_k \tag{1}$$

This assumption was empirically checked with 620 examinees who took three tests during a semester for a total of 117 items. At the .01 level of significance, it was found that all but 6 of the items satisfied this restriction (Wilcox, 1981b).

In Wilcox (1981a), it was proposed that the effectiveness of the distractors be measured with

$$H(q_1,\ldots,q_k) = -\sum_{i=1}^{k-1} p_i \ln p_i \tag{2}$$

where $p_i=q_{i+1}/(1-q_1)$. This is the entropy function which is also known as Shannon's measure of information or diversity. Wilcox (1981a) notes that if it is decided that an examinee knows the correct response if and only if the correct response is chosen on the first attempt of the item (i.e., a conventional scoring procedure is used) the distractors are the most effective when $q_2=q_3=\ldots=q_k$. This corresponds to random guessing, and Weitzman (1970) calls such items "ideal." The entropy function measures how far away an item is from being ideal. Small values of H indicate that guessing is not close to being random, while large values of H mean the item is close to being ideal. The largest possible value for H is ln(k-1), and its smallest value is zero.

For a random sample of n examinees, let x_i be the number who choose the correct response on the ith try. The maximum likelihood estimate of H is

$$\hat{H} = -\sum_{i=1}^{k-1} \frac{x_{i+1}}{n-x_1} \ln\frac{x_{i+1}}{n-x_1} \tag{3}$$

where the estimate is taken to be ln(k-1) when $n=x_1$ (cf. Gill & Joanes, 1979; Basharin, 1959; Hutcheson & Shenton, 1974).

The purpose of this note is to indicate how the polarization test recently proposed by Alam and Mitra (1981) might be extended to make inferences about H. Interest is focused upon testing the hypothesis

$$H_0: \ H < h$$

where h is a known constant. An important special case is h=ln(k-1) which corresponds to testing whether guessing is at random. The appeal of the procedure outlined here is that the exact distribution of a statistic used by Alam and Mitra, which is described below, can be used to compare H to h. This is important because asymptotic approximations of the distribution of H tend to be unsatisfactory unless n is very large (Bowman, et al., 1971). Comments by Alam and Mitra (1981) indirectly confirm this.

2. COMMENTS ON H, MAJORIZATION AND SCHUR FUNCTIONS

When making inferences about H, the natural procedure is to use $\hat{H}$ which is given in equation 3. However, the exact distribution of $\hat{H}$ is rather complex and cumbersome to work with (Bowman, et al., 1971). Instead the statistic

$$T(X) = \sum_{i=2}^{k} x_i^2 \qquad (4)$$

is used. Note that if T(X) is divided by $(n-x_1)^2$ we get an estimate of $\sum_{i=1}^{k} p_i^2$ which is known as Simpson's measure of diversity (Simpson, 1949).

At first glance it might appear that equation 4 is completely unjustified when making inferences about H, but in terms of majorization and Schur functions (which are defined below) this is not the case. The goal in this section is to briefly outline why this is true. Additional clarification of this point will be made in a later section.

Consider any two vectors $\underline{a}=(a_1, \ldots, a_k)$ and $\underline{b}=(b_1, \ldots, b_k)$, and let $a_{[1]} \geq a_{[2]} \geq \cdots \geq a_{[k]}$ be the components of $\underline{a}$ written in descending order. The vector $\underline{a}$ is said to majorize the vector $\underline{b}$, written $\underline{a} >^m \underline{b}$, or $\underline{b} <^m \underline{a}$, if

$$\sum_{i=1}^{j} a_{[i]} \geq \sum_{i=1}^{j} b_{[i]}, \qquad j=1, \ldots, k-1$$

and

$$\sum_{i=1}^{k} a_{[i]} = \sum_{i=1}^{k} b_{[i]}$$

where $b_{[i]}$ is defined in the same manner as $a_{[i]}$. For example,

$$(1, \ldots, 0) >^m (\tfrac{1}{2}, \tfrac{1}{2}, 0, \ldots, 0) >^m \ldots >^m (1/k, \ldots, 1/k).$$

A real valued function ϕ is said to be Schur convex if $\underline{a} >^m \underline{b}$ implies that $\phi(\underline{a}) \geq \phi(\underline{b})$. If $\underline{a} >^m \underline{b}$ implies $\phi(\underline{a}) \leq \phi(\underline{b})$, the function ϕ is Schur concave. In statistics there has been an increasing interest in Schur functions, and results in Alam and Mitra are formulated in terms of these concepts. For a recent summary of various results on Schur functions, see Marshall and Olkin (1979).

To motivate the use of equation 4, first we note that given x_1, T is a Schur convex function of $(x_2, \ldots, x_n)$, and H is Schur concave. This means that in the sense of majorization, both T and H can be used to measure the inequality of the p_i's, and indeed both measures are used. To put it another way, comparing H to h is comparable to comparing p to some known vector p_0, the comparison being made in terms of majorization. In fact, this is exactly what Alam and Mitra (1981) do in their paper, but they started with p_0 rather than h. As explained in more detail below, it is possible to start with h, and then formulate the problem in terms of comparing p to p_0. Once this is done, it is possible to make use of the results given by Alam and Mitra, but as will become evident, certain modifications of their results will be needed.

3. DETERMINING A p_0 WHEN h IS GIVEN

Suppose h has been specified. This section outlines how the problem of comparing H to h might be reformulated in terms of comparing p to some known vector p_0. First note that if $h=\ln(k-1)$, which is the maximum possible value of H(p); comparing H to h is the same as comparing p to $p_0=((k-1)^{-1}, \ldots, (k-1)^{-1})$.

Next let h be any real number between 0 and $\ln(k-1)$. Since H is Schur concave, there is an integer m such that

$$p_1=(1/m,\ldots,1/m,0,\ldots,0)>^m p_2=(1/(m+1),\ldots,1/(m+1),0,\ldots,0)$$

and $H(p_1)<h<H(p_2)$ where p_1 has m elements equal to m^{-1} and p_2 has m+1 elements equal to $(m+1)^{-1}$. Moreover, for any c such that $0<c<m^{-1}-(m+1)^{-1}$,

$$p_3=(m^{-1}-c, \ldots, m^{-1}-c, mc, 0, \ldots, 0)>^m p_2,$$

and $p_1>^m p_3$.

In addition, as c increases, p_3 decreases in the sense of majorization. Thus, for any h, $0<h<\ln(k-1)$, it is possible to find a vector p_0 such that $H(p_0)=h$.

For example, suppose an item has 4 alternatives. The maximum possible value for the entropy of the distractors is $\ln(3)=1.0986$. Suppose we want to determine whether the distractors have at least 80% of the maximum possible entropy. This corresponds to comparing H to h=.88. $H(\frac{1}{2},\frac{1}{2},0)=.693$, and so we determine p_0, $(\frac{1}{2},\frac{1}{2},0)>^m p_0>^m(1/3,1/3,1/3)$, such that $H(p_0)=.8$. For vectors of the form $(2^{-1}-c,\ 2^{-1}-c,\ 2c)$, c can be determined so that $H(\frac{1}{2}-c,\ \frac{1}{2}-c,\ 2c)=.8$. The answer is approximately c=1/32, and so $p_0=(15/32,\ 15/32,\ 2/32)$. In summary, comparing H to .88 is, in the sense of majorization, comparable to comparing p to $p_0=(15/32,\ 15/32,\ 2/32)$.

4. THE POLARIZATION TEST

The point of the previous section is that the problem of comparing H to h, or comparing any measure of diversity to a known constant, can be reformulated in terms of comparing an unknown vector to a known vector in the sense of majorization. This can be done if the measure of diversity is a Schur function. This section considers how p might be compared to p_0 once p_0 is determined.

The Distribution of T

The first step in devising a method of comparing $\underline{p}$ to $\underline{p}_0$ is to derive the exact distribution of T. First, however, we will need the distribution of

$$S(X) = \sum_{i=1}^{k} x_i^2.$$

where $X=(x_1, \ldots, x_k)$.

We note that expression (2.1) in Alam and Mitra (1981) is supposed to be the distribution of $S(X)$ for $k=2$. However, the maximum possible value of S is n^2, not n, and so the inequalities in their expression (2.1) are incorrect. Another problem is that the smallest possible value of S is $n^2/2$ if n is even, and $(n-1)(n+1)/2$ if n is odd; it is not $n/2$ as implied by Alam and Mitra's equation (2.1). The same mistakes are made in expression (2.2), and their expression (2.2) contains two other typographical errors. However, even if these corrections are made, the limits on the summation in their expression (2.1) are incorrect. As a simple example, suppose $n=3$. Then $\Pr\,(S(X)\leq 5)=\binom{3}{2}q_1^2(1-q_1)+\binom{3}{1}q_1(1-q_1)^2$ which does not agree with their results. Accordingly, the exact distribution of $S(X)$ is derived here.

First consider $k=2$, let $c_y=0$ if $y=n/2$; otherwise $c_y=1$. Let a be the smallest integer greater than or equal to $n/2$, and let b be the largest integer less than or equal to $z-n/2$, where z is the largest integer such that $z(n-z)\geq(n^2-s)/2$. Then

$$\Pr(S(X)\leq s) = \sum_{y=a}^{a+b}\left[\binom{n}{y}q_1^y(1-q_1)^{n-y}+c_y\binom{n}{n-y}q_1^{n-y}(1-q)^y\right] \qquad (6)$$

Next consider $k=3$. Since the joint probability function of x_2 and x_3 given x_1 is binomial with parameters $q_1/(1-q_3)$, $q_2/(1-q_3)$ and $n-x_1$,

$$\Pr(S(X)\leq s\,|\,x_1)=\sum_{y=a}^{a+b}\binom{n-x_1}{y}\left(\frac{q_2}{1-q_1}\right)^y\left(\frac{q_2}{1-q_1}\right)^{n-x_3-y}+c_y\binom{n-x_1}{n-x_1-y}\left(\frac{q_3}{1-q_1}\right)^{n-y-x_3}\left(\frac{q_2}{1-q_1}\right)^y \checkmark$$

where $n-x_1$ replaces n in the definition of c_y, a, b and z. Let $D_{k-1}(s,\ x_1)$ represent the right-hand side of this last equality where

$D_{k-1}(s, x_1)=1$ if $s \geq (n-x_1)^2$. If $n-x_1$ is even, $D_{k-1}(s, x_1)$ equals zero if $s<(n-x_1)^2/2$, and if n is odd it is zero if $s<(n-x_1-1)(n-x_1+1)/2$. It follows that

$$\Pr(S(X) \leq s) = \sum_{x_1=0}^{n} D_{k-1}(s,x_3) \binom{n}{x_1} q_1^{x_1} (1-q_1)^{n-x_1} \qquad (7)$$

For $k>3$, the distribution of $S(X)$ can be obtained recursively in the same manner.

Having established the distribution of $S(X)$, it is now possible to test the hypothesis $H>h$ which, via majorization, is comparable to testing $\underline{p} <^{m} \underline{p}_0$.

Let $B_k(s;q_1,\ldots,q_k,n)=\Pr(S(X) \leq s)$. In terms of the x_i's, the decision rule is to reject H_0 if

$$\sum_{i=2}^{k} x_i^2 > t.$$

where t is to be determined. From equation 7

$$\Pr\left(\sum_{i=2}^{k} x_i^2 < t \,|\, x_1\right) = B_{k-1}\,(t; p_1, \ldots, p_{k-1}, n-x_1) \qquad (8)$$

where, as before, $p_i=q_{i+1}/(1-q_1)$.

Since under the null hypothesis $p=p_0$, equation 8 can be evaluated, and so for any observed x_1, the probability of a type I error can be determined for any t.

An Illustration

As a simple illustration, suppose $k=3$ and it is decided to test

H_0: $H > .5$.

From section 3, p_0 is approximately (.8, .2) which, in the sense of majorization, is the same as (.2, .8). Thus, in terms of the polarization test,

$$H_0: \; p <^{m} (.8, .2).$$

Suppose n=100 examinees are randomly sampled and that $x_1=75$, $x_2=21$ and $x_3=4$ are observed. Then $\sum_{i=2}^{3} x_i^2=457$. Setting $p=(.8, .2)$, equation 8 yields the value of $Pr(T(X) \leq 425 | x_1=75)$ which in turn gives the value of $Pr(T(X)>425|x_1=75)$. Using the tables in Pearson (1968), the latter value was found to be .4206, and so the observed x's are reasonably consistent with the null hypothesis. If instead $x_2=24$, and $x_3=1$, $Pr(T(X) \geq 577|x_1=75)=.023$, and so the results would be significant at the .05 level.

An optimal property of the test. A desirable property of any hypothesis testing procedure is that as the unknown parameters move away from the null hypothesis, the power of the test increases. Here this means that if p' and p'' are any two vectors such that $p' >^{m} p'' >^{m} p_0$, we want the power of the test $p <^{m} p_0$ to be larger at $p=p'$ than it is at $p=p''$. That this property holds follows immediately from a theorem in Marshall and Olkin (1979, p. 391). Thus, we have an additional justifiction for using the polarization test as it is outlined above.

SUMMARY

In summary, the paper describes how hypotheses about the effectiveness of the distractors of multiple choice test items might be tested. Included as a special case is an exact test for random guessing that can be used in conjunction with an answer-until-correct scoring procedure. This is in contrast to the asymptotic test for random guessing (which does not use an answer-until-correct scoring rule) that was proposed by Weitzman (1970).

Another point is that it is not being recommended that an item be modified if H_0 is rejected. Wilcox (1981a) describes how the accuracy of a test item can be estimated. If the accuracy is high, there may be little reason for trying to improve the distractors by ensuring random guessing. The reason is that any improvements in the distractors might yield a negligible increase in item accuracy. However, if item accuracy is moderate or small, and if H_0 is rejected, consideration might be given to improving the distractors.

REFERENCES

Alam, K., & Mitra, A. Polarization test for the multinomial distribution. Journal of the American Statistical Association, 1981, 76, 107-109.

Basharin, G. On a statistical estimate for the entropy of a sequence of independent random variables. Theory of Probability and its Applications, 1959, 4. 333-336.

Bowman, K., Hutcheson, K., Odum, E., & Shenton, L. Comments on the distribution of indices of diversity. In G. Patil, E. Pielou, and W. Waters (Eds.). International Symposium on Statistical Ecology, Vol. 3, University Park: Pennsylvania State Press, 1979.

Gill, C., & Joanes, D. Bayesian estimation of Shannon's index of diversity. Biometrika, 1979, 66, 81-85.

Hutcheson, K., & Shenton, L. Some moments of an estimate of Shannon's measure of information. Communications in Statistics, 1974, 3, 89-94.

Marshall, A., & Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

Simpson, E. Measurement of diversity. Nature, 1949, 163, 688.

Weitzman, R. A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, in press (a).

Wilcox, R. R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, in press.

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1981, submitted for publication (b).